Estimating Traffic Signal States by Exploiting Sparse Low-Frequency Floating Car Data

Von der
Fakultät Architektur, Bauingenieurwesen und Umweltwissenschaften
der Technischen Universität Carolo-Wilhelmina
zu Braunschweig

zur Erlangung des Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte

Dissertation

von
Steffen Axer, M. Sc.
geboren am 25.02.1985
aus Henstedt-Ulzburg

Eingereicht am: 30.03.2017
Disputation am: 27.07.2017

Berichterstatter: Prof. Dr.-Ing Bernhard Friedrich
Prof. Dr.-Ing Markus Friedrich

(2017)

Danksagung

Ich möchte mich in erster Linie bei Prof. Dr.-Ing. Bernhard Friedrich für die Betreuung dieser Doktorarbeit bedanken. Die äußerst konstruktiven Gespräche und Ratschläge sowie die hervorragende technische Ausstattung des Instituts haben einen essentiellen Beitrag zur Entwicklung dieser Arbeit geleistet. Zudem bedanke ich mich herzlich bei Prof. Dr.-Ing. Markus Friedrich für die Übernahme des Koreferats.

Weiterhin möchte ich mich insbesondere bei meinen ehemaligen Kollegen Ralf Kutzner, Federico Pascucci und Sebastian Vogt für die vielen fruchtbaren Diskussionen sowie für den Aufbau und die Versorgung diverser Simulationsszenarien bedanken.

Ein ganz besonderer Dank geht an meine Familie sowie an meine Ehefrau Caroline, die mich während der gesamten Entstehungszeit dieser Arbeit mit all ihrer Kraft begleitet, motiviert und unterstützt hat. Ohne euch wäre diese Arbeit nicht möglich gewesen.

Abschließend bedanke ich mich bei der Abteilung Entwicklung Verkehrsinformation des ADAC für die Bereitstellung notwendiger Floating Car Data sowie der Abteilung Verkehrsmanagement des Kreisverwaltungsreferates der Stadt München für das zur Verfügung stellen notwendiger Referenzschaltzeitdaten.

Kurzfassung

Zukünftige Fahrerassistenzsysteme benötigen häufig netzweit verfügbare Informationen über den zu erwartenden Signalzustand an einer Lichtsignalanlage (LSA). Aktuelle Verfahren zur Signalzustandsschätzung basieren dabei auf der Auswertung von Zustandsdaten einzelner LSA, die über eine zentrale Servereinheit gesteuert werden. Die flächendeckende Versorgung von Signalzustandsschätzungen ist somit ein aufwändiges Unterfangen, da Datenverbindungen zu zahlreiche Verkehrsbehörden aufgebaut werden müssen. Die zunehmende Verbreitung Smartphone-gestützter Navigationssysteme sowie vernetzter Fahrzeuge hat in den letzten Jahren zu einem starken Anstieg sogenannter Floating Car Data (FCD) geführt. Die nahezu flächendeckend verfügbaren Daten ermöglichen die effiziente und netzweite Schätzung verkehrsrelevanter Informationsgrößen, wie z.B. Signalzustandsschätzung von LSA.

In dieser Arbeit wurde eine neuartige Methode zur Schätzung von Signalzuständen an LSA auf Basis realer FCD entwickelt und unter Realbedingungen getestet. Im Vorfeld dieser Entwicklung wurde der aktuelle Forschungsstand zu diesem Themenfeld analysiert. Um die Datenbeschaffenheit einer realen, kommerziell verfügbaren FCD-Quelle zu verstehen, wurde im Zuge der Arbeit ein jährlicher FC-Datensatz des ADAC (Allgemeiner Deutscher Automobil-Club) an zwei ausgewählten Knotenpunkten in München analysiert.

Motiviert durch das verhältnismäßig geringe Trajektorienaufkommen der genutzten Datenquelle, wurde eine durchgängige Methodik zur Verarbeitung weniger Trajektorien mit zeitlich niedriger Datenauflösung entwickelt. Die Methodik ermöglicht dabei die Schätzung werktäglicher, typischer Signalprogrammparameter, wie z.B. der Umlauf- und Freigabezeit für einzelne Stunden im Tagesgang. Die berechneten Parameter können unter definierten Randbedingungen zur Schätzung zukünftiger Signalzustände verwendet werden.

Zur möglichst effizienten Datennutzung wurde ein neuartiges, auf einem Hypothesentest basiertes Verfahren zur Schätzung der Umlaufzeit entwickelt. Die Methodik wertet historische Querungsvorgänge an der Haltlinie einer

LSA aus. Die beobachtenden Ereignisse werden mittels Modulo Operator in den Zeitstrahl einer hypothetischen Umlaufzeit projiziert. Ausschließlich die Verwendung der korrekten Umlaufzeit führt zu einer Konzentration der Querungsvorgänge innerhalb der Umlaufzeit. Dieser Effekt kann mittels Hypothesentest statistisch zur Identifikation der wahren Umlaufzeit ausgenutzt werden. Das entwickelte Verfahren wurde mit einer Monte-Carlo-Simulation analysiert.

Ein weiteres wesentliches Ziel der Arbeit war die robuste Schätzung von Freigabezeiten an einzelnen Fahrströmen unter Nutzung der zur Verfügung gestellten Daten. Das hierzu neu entwickelte Verfahren wertet Querungsvoränge sowie die Anzahl wartender Fahrzeuge über die Umlaufzeit aus. Mittels eines iterativen Klassifikationsprozesses werden einzelne Umlaufsekunden in Abhängigkeit des Datenaufkommens zu einer zeitlich konsistenten Freigabezeit verarbeitet. Hierzu berechnet das Verfahren einen selbst-kalibrierten Häufigkeitsgrenzwert, den es gilt zu überschreiten. Zur Plausibilisierung des Freigabezeitintervalls setzt die Methodik auf einen ebenfalls iterativen Selbstprüfungsprozess, bei dem es gilt, daten- und zeitspezifische Randbedingungen einzuhalten. Die Prüfmethodik gewährleistet robuste Ergebnisse, auch bei einem geringen Trajektorienaufkommen.

Darüber hinaus widmet sich die Arbeit mit der Entwicklung eines Verfahrens zur Freigabezeitüberprüfung an Konfliktströmen einem bisher noch nicht ausreichend betrachteten Forschungsthema. Konflikte zwischen geschätzten Freigabezeiten treten dabei insbesondere an verkehrsabhängigen LSA auf. Wesentliche Voraussetzung für die entwickelte Methodik ist das Vorhandensein von Freigabezeitschätzungen auf allen Konfliktströmen des Knotenpunktes. Zur Erreichung dieser Anforderung wurde ein Clusterverfahren zur Erkennung von Stunden mit ähnlichen Signalprogrammparameter entwickelt. Die gruppierten Stunden können zusätzlich aggregiert werden und ermöglichen damit auch die Berechnung von Freigabezeiten auf Fahrströmen mit geringem FCD Aufkommen. Der Ansatz wurde im Zuge einer Simulationsstudie hinsichtlich der korrekten Funktionsweise überprüft.

Die Methodik zur Schätzung von Signalprogrammparametern sowie zur Prädiktion von Signalzuständen wurde unter Realbedingungen an vier ausgewählten LSA in München kalibriert und validiert. Die erreichte Qualität der Zustandsschätzung wurde mittels unterschiedlicher statistischer Indikatoren analysiert und bewertet.

Abstract

Many emerging driving assistance concepts require a network wide accessible information of future traffic signal states. Nowadays traffic signal state prediction approaches are strictly limited to work only at individual cities, as these methods require a data interconnection to a centralized traffic signal control server. The increasing availability of Floating Car Data (FCD) caused by the popular use of smartphone navigation and connected vehicles offers a very promising data source that allows to deduce cost-efficient estimates of traffic related information for large networks, such as signal timing parameters and traffic signal states.

In this thesis, a novel methodology for the traffic signal state prediction based on real-world FCD has been developed and evaluated. In advance of this development process, the thesis presented the latest state-of-the-art of a trajectory-based traffic signal state prediction. In order to understand the data characteristics of a real-world, commercially available FCD source, this work analyzed a one year dataset obtained from the ADAC (General German Automobile Club) at two major intersections in Munich (Germany).

Motivated by the obtained, sparse and likewise low-frequently sampled trajectory volumes, this work developed a comprehensive methodology that covers all major aspects of the trajectory processing in order to infer hourly estimates of the signal timing parameters, such as the cycle length and the green intervals typically used on workdays. The signal timing parameters allow moreover the prediction of future traffic signal states with respect to different described constraints.

To utilize the trajectory volumes efficiently, the thesis developed a novel, hypothesis-based cycle length estimation procedure. The approach estimates the typical cycle length for an hour by analyzing vehicle stop bar crossing events. These events are projected into the time scale of potential cycle length candidates by the help of a modulo operation. Only the correct cycle length causes a concentration of crossing events within the tested cycle time. To infer the true cycle length, concentrated trajectories are exploited by the help of

a hypothesis test. The developed concept has been analyzed for a fixed-time controlled intersection by means of a Monte-Carlo-Simulation.

The next major objective of this thesis was the development of a robust green interval estimation methodology that is capable to deduce the green interval by exploiting sparse crossing events as well as stopped vehicles. The novel concept is based on an iterative classification procedure that infers the green interval by analyzing the number of counted crossing events per cycle second. The approach determines automatically a reasonable data threshold that needs to be exceeded in order to classify a cycle second to the signal state green. A rule-based and iterative self-verification process checks the consistency of the found green interval, which makes the methodology robust when working with small trajectory samples sizes.

In addition, the research of this thesis closed the gap of a missing green interval verification procedure that checks potentially overlapping green intervals of conflicting movements, mainly caused by traffic actuation. As the developed concept requires ideally green interval estimates on all conflicting movements, the thesis developed moreover a novel cluster-based approach that tries to identify hours with very similar signal timing parameters. The trajectory data set is once again aggregated over the found time cluster to enable even on movements with less volumes the estimation of signal timing parameters. The correct functionality of the developed clustering procedure was tested with the help of simulation study.

Finally, the developed traffic signal state prediction concept was calibrated and validated at four real-world intersections in Munich (Germany), accompanied by a set statistical indicators in order to assess and analyze the traffic signal state prediction quality.

Contents

CHAPTER 1

Introduction

1.1 Motivation

The worldwide increasing degree of urbanization [73] is a major key factor that influences massively people's quality of living. On the one hand, economists could state that the agglomeration of populations in urban areas in generally will increase the output per capita, which causes in tendency better living standards for our society. On the other hand, traffic engineers would state that rapidly growing urban areas could cause likewise massively negative effects such as environmental degradation [12], if the transportation systems and urban infrastructure is not permanently adapted to manage the increasing amount of people.

Referring to a recent prognosis [73], in the year 2030 approximately 78 % of the Germany population will live in highly urbanized areas. As a consequence of this trend, politics tighten gradually European emission standards to urge infrastructure operators as well as the automotive companies to create a smarter and more efficient mobility. Different research has shown that especially traffic lights [29, 58] have a significant influence on the road network capacity and can impact intensively the emissions of motorized traffic. In further consequence, traffic engineers developed smarter signal control strategies that try to reduce delays and vehicle stops by optimizing traffic lights' parameterization within the road network. Moreover, automotive companies also contributed with the help of various traffic light assistance concepts that allow drivers as well as the vehicle to be well adapted to future traffic signal states. Thus, various traffic light assistance concepts are in general imaginable that range from a Green Light Optimal Speed Advisory (GLOSA) over an Adaptive Route Change (ARC) to a simple Red Light Duration Advisory (RLDA) [42, 39]. However, all of the mentioned assistance concepts require the knowledge of future traffic signal states, whereas necessary accuracy depends highly on the specific use case and the preferred human-machine interface.

To be able to predict traffic signal states, most methodologies [14, 10] focus on the processing of traffic signal switching and vehicle detector data, supplied by a centralized traffic control server. In order to establish the desired assistance functionality, forecasted signal states are typically transmitted via short range communication technologies [38] or mobile phone networks to the vehicle. The main concerns of this concept are twofold. First, the methods require a permanent data interconnection among all traffic lights within the territory of the road authority in charge. Second, a service provider that has the willingness to establish a national or even international service needs to request data access for every single road authority and handles probably different types of data interfaces, which could become a financial challenge.

During the recent years, the increasing availability of Floating Car Data (FCD) caused by the popular use of smartphone devices and connected vehicles offers a very promising data source that has already revolutionized the way how traffic engineers can nowadays access network-wide and cost-efficient traffic related measures, like travel times and speeds [68, 4, 9]. To overcome the problem of a spatially limited availability of traffic signal state predictions, this thesis develops and evaluates a novel concept that exploits sparse volumes of low-frequently sampled FCD trajectories to estimate signal timing parameters like cycle length and green intervals at individual real-world intersections. The estimated signal timing parameters are moreover used to predict future traffic signal states. Subsequent Section 1.2 provides a dedicated overview about all intended objectives.

1.2 Objectives

The research of this thesis has been initially motivated by the work of Kerper et al. [40] who explored already in 2012 the general feasibility to estimate signal timing parameters and to predict traffic signal states at fixed-time controlled intersections by exploiting trajectory data. Many important questions remained, because the authors utilized only a simulated dataset, in which trajectories have been sampled second-by-second without considering typical real-world effects, like positioning inaccuracy and low trajectory volumes.

Since 2012 other researchers such as Schönauer et al. [72], Axer et al. [8], Fayazi et al. [26] and Protschky et al. [62] have contributed to the topic in a short time period and a variety of different datasets and methods have been used to estimate traffic signal states. Thus, an important objective of this thesis is to investigate the available concepts and to discuss their advantages and disadvantages. This investigation showed clearly that especially the analysis of real-world FCD with sparse volumes and likewise low-frequently sampled trajectories have not been explicitly analyzed in the past. Therefore, research of this thesis has the objective to intensify the work within this special research area.

The main objective of this thesis is the development of a comprehensive methodology that is capable to operate on real-world trajectory data, supplied by a single commercial FCD provider. The method should allow an estimation of signal timing parameters which are used to predict future traffic signal states at fixed-time and even traffic actuated intersections. The research thereby focuses especially on the required data processing as well as the achievable estimation quality. The application of derived estimates in different use cases is not within the scope of this thesis.

To allow the network-wide analysis of trajectories with such properties, this work has moreover the objective to investigate and describe comprehensively all relevant preprocessing tasks such as the road graph provisioning as well as the map matching and path inference, needed to project real-world vehicle trajectories on a road graph model. The developed and implemented preprocessing algorithms are the basis to allow the creation of an estimation methodology that is able to operate on sparse real-world trajectory volumes. As these volumes are in contrast to other research [40, 72, 26, 62] very small, this thesis focuses especially on a robust estimation concept that allows to prove the consistency of the trajectory data with respect to various predefined rules.

An other important scope of this thesis is the self-verification of signal timing estimates at an intersection by comparing and adjusting the estimations of conflicting movements in order to derive results that are consistent with the required intergreen times.

The very last objective of this thesis is the establishment of a very thorough calibration and validation methodology in which estimated signal timing parameters are used to predict traffic signal states at different real-world intersections. The thesis has thus the scope to present transparently the estimation quality at real-world intersections with respect to different statistical measures.

1.3 Outline

Chapter 2 provides an overview of the latest state-of-the-art of science and technology. Basic fundamental terms and methods of traffic signal control will be presented followed by brief introduction of the FCD concept. In addition, the chapter reviews related research and provides a summary of need for further research, addressed in this thesis.

Chapter 3 presents preliminary a real-world FCD trajectory volume analysis in order to understand the data situation of a commercially available FCD source, used in this thesis. Based on this work, the fundamental concept and major constraints of the developed signal timing estimation methodology will be introduced. Moreover, the chapter describes the fundamentals of statistical indicators, used in the later real-world calibration and validation procedure.

Chapter 4 deals with the data processing and the estimation methodologies, developed in this thesis. Preliminary preprocessing steps such as the map matching will be introduced followed by an explanation of the exploited trajectory information. In addition, this chapter introduces step by step the procedure of the cycle length and the green interval estimation. Moreover, a result clustering and verification procedures is introduced. All concepts are accompanied by different simulation studies.

Chapter 5 describes a detailed real-world calibration and validation procedure. Further on, important results and implications are discussed in order to identify potentials for improvements.

Chapter 6 closes with a summary and an outlook on further research.

CHAPTER 2

State-of-the-art of Science and Technology

2.1 Basics of Traffic Signal Control

Traffic signal control is a major factor of urban traffic control and management which is mainly used for reasons of traffic safety and flow regulation. Thus, the principle of signal control could be applied for single isolated intersections, a road stretch or even for complex meshed networks covering multiple intersections. As Section 2.2 has the objective to introduce briefly aspects of traffic signal control strategies, this paragraph clarifies some fundamental terms, frequently used in this thesis, whereby definitions refer to Pohlmann's dissertation [60]:

- **Signal head:** a single traffic signal that can display different signal indications such as green, red, or the intermediate signal indications amber and/or red-amber.
- **Signal group:** a set of one or more traffic signal heads that display at all times the same signal indication for a single turning movement or lane respectively (left, through, right), or for a combination of these turning movements on separate or mixed lanes.
- **Signal plan:** a set of switching times that indicates the points in time in seconds when each signal group changes its signal indication from green to red and vice versa (fixed time signal plan).
- **Phase:** a time interval of a signal plan during which the signal indication of all signal groups remains unchanged. Green times of individual signal groups may start and end at different times.
- **Intergreen time:** the time interval between the end of the green signal indication of a signal group and the start of the green signal indication

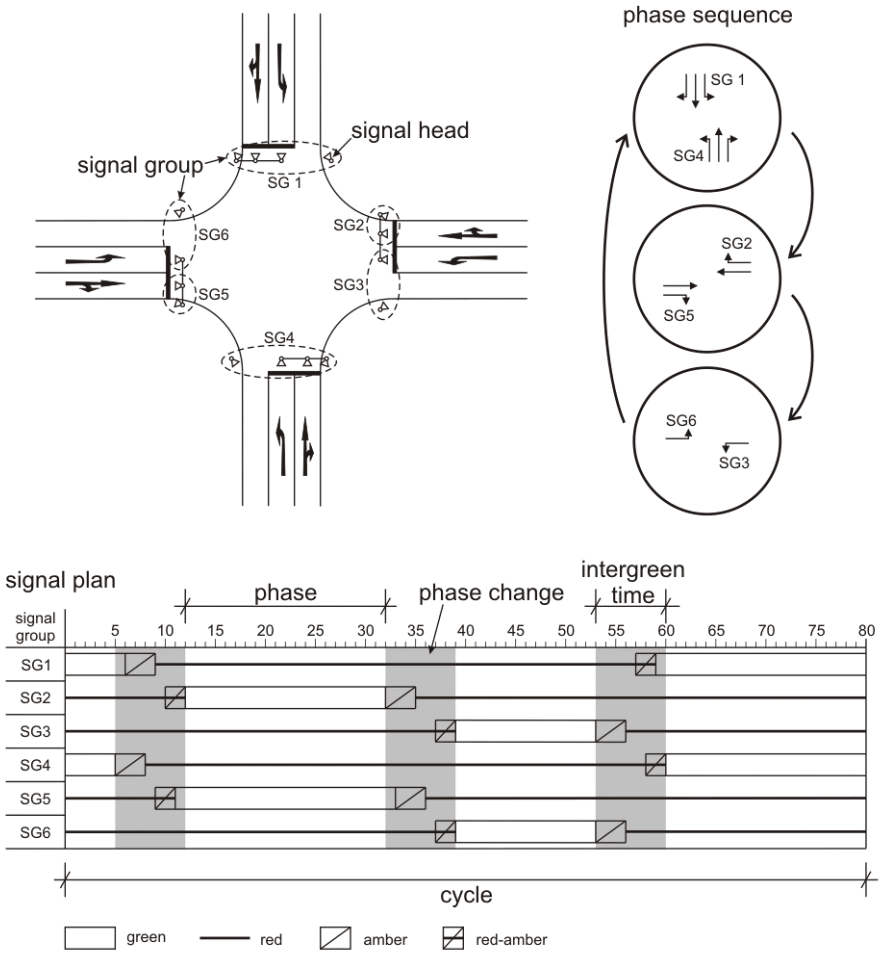


Figure 2.1: Fundamental terms of traffic signal control, taken from Pohlmann [60]

of a conflicting signal group. This time interval must be long enough to allow clearing vehicles to pass the conflict area safely before entering vehicles reach it.

- **Phase change:** a time interval of a signal plan during which the signal indication of all affected signal groups changes from one phase to the next. The phase change contains intermediate signal indications for all signal groups whose signal indication changes between the two phases. It starts at the end of the green time of the earliest changing signal group and ends at the beginning of the green time of the latest changing signal group.
- **Cycle:** a single repetition of a signal plan consisting of a sequence of phases and phase changes such that each signal group displays a green signal at least once per cycle. Each cycle has certain duration which is defined by the cycle length.

- **Phase sequence:** the sequence of phases that is applied during one cycle.
- **Offset:** an amount of time in seconds by which the entire signal plan of an intersection is shifted in time against a reference point in time. This concept is called the absolute offset. The offset can take a value between zero and the cycle length minus one. It is used at arterials or in networks to coordinate signalized intersections in a progressive signal system ("green wave") such that (ideally) platoons of vehicles can pass a sequence of intersections in a so called green band without stopping. The difference of the absolute offsets of two adjacent intersections is called the relative offset between these intersections. It is equal to the time span between the start of the green signal indications at both intersections for the coordinated traffic stream.

2.2 Traffic Signal Control Strategies

According to the German "Richtlinien für Lichtsignalanlagen - RiLSA" (Guidelines for Traffic Signals) [28] traffic control strategies can be differentiated with respect to a macro- and microscopic control level. The macroscopic control level reacts thereby on long-term historical traffic measures like typical traffic demand, densities or tailback lengths. This more strategical control level is used to select preconstructed general framework signal plans, whereas the selection process can consider the time-of-day or general traffic demand situation, deduced on strategical vehicle detectors within the road network. In contrast thereto, the microscopic control level aims on the objective to react more flexibly to local short-term variations of the traffic demand with the help of different control strategies, summarized in the following paragraphs.

2.2.1 Fixed-Time Control

The simplest strategy for traffic control are fixed- or pre-timed controlled intersections. The strategy considers thereby a constant phase sequence, phase changes and green intervals for a certain signal plan. Parameters are typically determined for mean traffic demand scenarios, deduced from historical traffic volumes. Fixed offsets between consecutive intersections could be established to allow vehicles passing intersections in platoons by assuming a fixed cruising speed. As this control strategy is not able to react on short-term stochastic traffic demand variations, a set of different pre-calculated signal plans is mostly used on the macroscopic control level that fits to typical morning, afternoon and off-peak traffic demand periods. Used signal plans are stored and selected by the signal controller with respect to the time-of-day or measured traffic volumes, deduced from strategic detectors within the road network [60].

2.2.2 Vehicle Actuated Control

In contrast to fixed-time control, vehicle actuated control strategy has been developed to react on local short-term stochastic variations of traffic demand. One or multiple approaching lanes are equipped with vehicle detectors that measure volumes, headways and occupancies. Gathered measurements enable the controller to trigger variations in the frame signal plan. Compared to a fixed-time signal plan, green intervals are allowed to vary within a certain time range. Traffic engineers design typically overlapping periods within the cycle time during which different phases may be displayed. This technique enables the traffic signal controller to decide when to start and end each phase with respect to the measured demand. Depending on the designed degree of actuation, phases can be aborted, extended, skipped, inserted and swapped which can lead to very different phase durations and sequences during each cycle. In order to achieve a desired degree of traffic actuation, processing logics are constructed that need to meet legal constrains of minimum and maximum green and red intervals. Because actuated control strategy is only applied for isolated intersections, coordinated green waves can be established by the help of fixed offsets that shift the framework signal plans for adjacent intersections. Research has shown that especially during periods of high traffic demand which are close to the intersection's capacity, vehicle actuated control tends to act like fixed-time control since all phases are extended to their maximum phase duration [77, 60].

2.2.3 Adaptive Traffic Control Systems

Previously described signal control strategy allows only to react to stochastic changes of traffic flow for single intersection with the help of a pre-designed processing logic. In contrast thereto, the concept of Adaptive Traffic Control Systems (ATCS) has the objective to optimize traffic signal control for much larger parts of a road network by considering current and future traffic demand situations. ATCS exploits typically multiple vehicle detectors within the road network to estimate current and upcoming traffic demand in order to infer signal timing parameters for all intersections that fit well to the current and future traffic demand situations. The knowledge of traffic demand is utilized to simulate traffic as well as environmental impacts for sets of signal timing parameter with respect to a certain time horizon. Most ATCS follow thereby the objective to minimize or maximize a certain performance index like total delay, number of stops and others by varying signal timing parameters like offsets or cycle length. The set of parameters that produces the best performance index is chosen to be used during the next time interval [60].

2.3 Technical Concept of Floating Car Data

Having explained the fundamentals of traffic signal control, this section introduces the technical concept of Floating Car Data (FCD). FCD describes a method in which traffic related information is gathered by the help of vehicles that act like a mobile sensor moving or floating through the road network and communicating collected information to a centralized server infrastructure. In order to determine vehicles' location within the road network and to communicate sensed information to the infrastructure, vehicles are equipped with a positioning device, such as GPS (Global Positioning System) and a cell phone based communication device [70].

Server-sided stored information describes for each vehicle a sequence of consecutive data points $p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_n$ whereas each data point stores at least a 3-tuple $p = (x, y, t)$ containing spatio-temporal information. The time interval between adjacent data points is the sampling interval t_d, typically integer multiples of a second. The reciprocal sampling interval $1/t_d$ describes the sampling frequency whereas the term "low frequency" implies commonly the usage of sampling intervals from 10 to 120 secs [19]. A consecutive sequence of data points is synonym to a spatio-temporal vehicle trajectory T which contains additionally very often the instantaneous driving speed, measured by the GPS device. Such a simple trajectory could be extended by further attributes, so called xFCD, measured by various vehicle sensors to acquire environmental or motion specific information [36].

In contrast to FCD, Floating Phone Data (FPD) describes a technical concept that exploits location-area-updates recorded from any smartphone in standby-mode and allows thus to reconstruct user's path through the road network. On the one hand, FPD allows in comparison to FCD the collection of higher trajectory volumes reasoned by good penetration rates. On the other hand, location accuracy is clearly inferior to FCD, as position could be only estimated roughly with respect to the cell phone network structure [71].

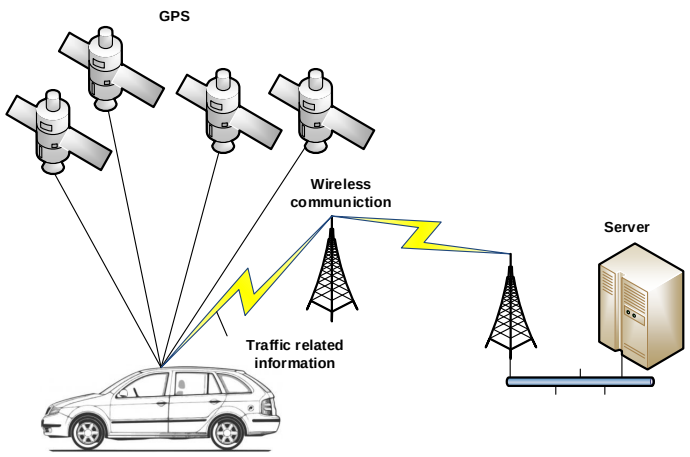


Figure 2.2: Technical concept of FCD, based on Schäfer et al. [70]

To establish traffic related services, a data processing is needed that matches the vehicle trajectory with the underling road graph (see Section 4.1.2). Map matched trajectories can be aggregated over different time horizons to estimate information like typical speeds, travel times or origin-destination matrices within our road networks. Given a sufficient FCD penetration, the collected data is capable to estimate network-wide traffic related measures which made FCD in recent past much more valuable. The increasing FCD penetration could be explained by the popularity of GPS enabled smartphones that allow online navigation and a high variety of location based services which require knowledge of user's position. In addition, an increasing number of vehicles is nowadays by default equipped with GPS and wireless communication devices that allow automotive companies to collect anonymized large volumes of FCD. Till now, the total number of global player such as TomTom, Inrix, GARMIN, Google, HERE that are collecting commercially large scale FCD volumes over multiple western countries is relative limited whereas collected data is hardly accessible for public research. Beside these global players, smaller national or locally operating companies like taxi services are often used in research to evaluate new concepts of urban traffic control and management.

Thus, this thesis exploits a FCD source obtained from the ADAC (General German Automobile Club), an automobile club in Germany that utilizes FCD to provide members and public users traffic related services [1]. As fleet size and origin of collected vehicles is not revealed to the author, Section 3.1 has the objective to clarify specifics like available trajectory volumes and sampling intervals in order to make research comprehensible.

2.4 FCD-based Traffic Signal State Prediction

Motivated by the advantages of an improving and network-wide available data source, different research has been carried out in the more recent past that tries to exploit trajectory data to estimate signal timing parameters and predict traffic signal states. Developed methods incorporate very different data requirements and principles, which will be introduced in the following paragraphs.

In 2012, research started with Kerper et al. [40] by investigating the feasibility of signal timing estimation at fixed-time controlled intersections by exploiting only simulated trajectories without considering positioning or timing errors. The developed data analysis utilized artificially generated trajectories with a sampling interval of $t_d = 1$ sec and a relatively high FCD penetration rate of approximately 5 %. The established method requires knowledge of stop bar crossing events and stopped vehicles. Trajectories that recorded both situations are used to deduce the start of green whereas all measured green starts are projected with the help of a modulo operation into the time scale of a

cycle length candidate. Kerper et al. [40] are iterating over a range of potential candidates in order to find a reasonable cycle length estimator that minimizes mean square error between projected consecutive green starts. In a next step, green and red intervals are deduced cycle-per-cycle by counting crossing and stopped vehicles within the cycle time. Analyzing sequentially short historical time periods of some cycles allows to produce a time series view of the inferred signal timing parameters which can be used to predict future signal states. Authors proved in simulation that under very simplified conditions signal states could be estimated for fixed-time controlled intersections without any prediction error.

Research from Kerper et al. is a very important contribution because they exploited the periodicity of fixed-timed controlled intersection to deduce signal timing parameters for the very first time. Nevertheless, the developed methodology could be only validated at simulated fixed-time controlled intersections, which means the approach has never been proved under typical harsh real world conditions in which positioning errors and much higher sampling intervals need to be managed.

In 2013, Krijger [43] contributed to the research topic with his master thesis. The fundamental principle of his approach assumes that historically recorded and aggregated crossing events on conflicting through movements of major and minor roads should be clearly separated if one projects them by the help of a modulo operation into a potential cycle length. An algorithm iterates over a range of potential cycle length candidates in order to find a certain cycle length which minimizes the overlap of crossing events on conflicting through movements. Furthermore, his approach assumes that only two vehicles are enough to estimate in principle at fixed-time controlled intersections the total duration of the green interval on one through movement, if these vehicles occurred in the same cycle and they could sense the start and end of green interval. Most of his very detailed and statistical analysis has been carried out with a simulated trajectory dataset in which Krijger modeled the vehicle arrival process with the help of a Poisson-distribution. His thesis concentrates thereby on the needed observation time to estimate signal timing parameters at fixed-time controlled intersection within a desired quality by testing a large bandwidth of FCD penetration rates that range between 0-50 %. A very short paragraph of Krijger's thesis is dedicated to the concept validation at a single fixed-time controlled intersection in Portland (USA) and an actuated intersection in Helmond (Netherlands) by using real trajectories, provided by TomTom. Caused by the high degree of actuation, validation in Helmond was not successful, whereas signal timing parameters at fixed-timed controlled intersection in Portland could be estimated and used temporarily for a traffic signal state prediction. Statements and statistical evaluations about the prediction errors are very sparse. Thus, Krijger stated that his method was

able to predict the green start on two movements with an absolute error of two respectively three seconds.

The most important contribution of Krijger's thesis is the innovative idea to estimate the cycle length by comparing crossing events of conflicting through movements. All in all Krijger has proved that the idea of signal timing parameter estimation based on trajectories has a great potential which could not be exploited optimally with his method due to a lack of trajectories. This means that his approach was not well adapted to operate with very small trajectory volumes.

In 2014, Schönauer et al. [72] published a purely data-driven approach to estimate signal timing parameters and predict traffic signal states at fixed-time controlled intersections with the help of bicycle trajectories. In contrast to all other approaches, their method tries to infer the correct cycle length by analyzing waiting periods of cyclists in front of intersection's stop bar. The fundamental concept of cycle length inference relies on the fact that cyclists arriving during a green interval should have no waiting times while cyclists arriving during a red interval are forced to stop and wait for a certain period. Schönauer et al. are measuring the waiting period of trajectories and project them with the help of a modulo operation to potential cycle length. They learned empirically that waiting times, projected to the correct cycle length, form a highly concentrated and coherent pattern, visualized in a second-wise histogram. The histogram abscissa describes thereby the cycle time whereas ordinate counts the frequency of stopped cyclists per cycle second. They exploit this effect by searching systematically a certain cycle length at which the histogram reveals a maximum of empty bins, whereas a bin is assumed to be empty, if frequency remains under a certain frequency threshold. In a next stage, Schönauer et al. try to infer the start of green by analyzing frequency gradients in the histogram. Thus, the cycle second which reveals the highest drop of stopped cyclists is used as the green start. Start of the red interval is obtained by cumulating stopped cyclists per cycle second backward, till the 99th percentile is reached.

All in all, the method worked very well at fixed-timed controlled intersections in Vienna (Austria) and the authors stated that the absolute signal state prediction error is approximately 5 secs. Nevertheless, the researchers had very good data conditions which could probably not be reached in the near future, when working with commercial vehicle-based FCD. The trajectories have been obtained manually with customized GPS receivers by a group of 20 cyclists. Collected and filtered trajectory volumes supplied the analysis process in minimum 100 qualified trajectories, whereas position information has been logged second-by-second. The used concept of a histogram-based data analysis can be concluded to be well performing. However, the authors did not need to consider smaller trajectory volumes and larger sampling intervals. It is obvious that a smaller number of stopped cyclists can cause

a volatile histogram that makes gradient analysis of frequencies unreliable. Furthermore, the authors did not explain how to determine the frequency threshold which needs to be exceeded in order to classify a cycle second to the signal state red.

In 2015, the research group around Fayazi et al. [26] continued the work from Kerper et al. [40] by using real-world trajectories with a sampling interval that ranges from 10 to 80 secs. Highly recurrent trajectories have been supplied by a bus transit system in San Francisco. The research study is based on data of two frequently used bus routes which allowed the collection of 4289 trajectories per month. Fayazi et al. developed a multi-staged concept that filters in a first instance trajectories that contain the driving sequence of deceleration, stop and acceleration. Filtered and aggregated trajectories allow thus the inference of the red interval for single movements. In the second stage, qualified trajectories are again exploited to infer single green start events by the help of a simple driving dynamic model that tries to calculate the time point at which a vehicle started its acceleration. In a third stage, the cycle length is deduced similarly to Kerper et al. [40] by applying a modulo operation on consecutive measured green starts. The method assumes a fixed-time signal control and tests a range of potential cycle length candidates by projecting all absolute green start events into the time scale of cycle length candidates. The deviation between projected green start events is analyzed in order to find a cycle length estimator which shows a minimal standard deviation. All three attributes, green start, red interval and cycle length are then observed in a time series analysis to identify different signal plans for fixed-time controlled intersections on the bus route in San Francisco. The authors validated their method more detailed at a single fixed-timed controlled intersection at which one was able to predict the green start with a root mean square error of approximately 3 secs.

Research of Fayazi et al. is a valuable contribution because they developed and validated a method that is capable to process real world trajectory data with a high sampling interval. The main advantage of the introduced approach lies similar to Kerper et al. [40] in the ability to estimate signal timing parameters in a time series which enables in principle the detection of a changed signalization, such as modified cycle lengths and green intervals. Nevertheless, the method exploited a very specialized trajectory dataset that provided relative high volumes. Thus, the high periodicity of buses allowed the collection and examination of an extraordinary number of trajectories in a very short time period. The weak spot in the approaches from Kerper and Fayazi is the exploitation of a trajectory subset that recorded subsequent green starts to estimate the cycle length. Thus, their own comment in [26] makes clear that the cycle length inference fails, if the movement of interest is a part of a well coordinated green wave at which only a very small number of trajectories are forced to stop.

Likewise in 2015, Protschky et al. [62, 63] published two different approaches to estimate the cycle length. Their first method [62] refers to Krijger [43], who exploited already the principle that crossing events on conflicting through movements should not overlap each other. The main drawback of this approach is the general requirement to have enough trajectories on both movements. If on the minor road less or no trajectories could be sampled, the method fails. Protschky et al. [62] focused on a statistical analysis to figure out the required trajectory volume to infer the true cycle length. They concluded that at intersections, served with two signal phases, approximately 40 trajectories per conflicting through movement are sufficient enough to estimate the cycle length with a confidence of 99 %. An increasing number of signal phases could require up to 100 trajectories per conflicting movement.

Protschky et al.'s second approach [63] is capable to infer the cycle length for individual movements, which is a great improvement. Therefore, a combination of three statistical measures is used, ranging from a variance analysis of stop points, trajectory correlation up to image analysis. The proposed method has its strengths and Protschky et al. could state that the approach worked for 39 real-world, fixed-time controlled intersections. They figured out that approximately 30 trajectories per movement were sufficient enough to estimate the cycle length of fixed-time controlled intersections with a confidence of 90 %.

The method shows a very good performance on certain, preselected fixed-time controlled intersections at which trajectories formed typically very regular patterns. However, the authors could not evaluate the approach on traffic actuated intersections at which perhaps not all of the required three statistical measures can be calculated, reasoned by a much stronger disturbance of the regular trajectory pattern. Furthermore, the authors did not clarify the used trajectory dataset and its specifications. The presented, very smooth trajectory plots indicate that the used dataset needed to have a very small sampling interval of approximately one second or lower [63]. Beside the cycle length estimation, Protschky et al. [63] proposed also a method to estimate the green interval by exploiting again stop bar crossing events. The method assumes that over the cycle time normalized empirical crossing events can be in general modeled by an inverse Gaussian distribution. The authors proved this hypothesis only visually at a single fixed-time controlled intersection in Munich (Germany). Knowing the cycle length, the method calculates the probability of a successfully observed crossing event at the beginning and the ending of an assumed inverse Gaussian distribution by testing different sets of distribution parameters. The distribution set which seems to fit best to the empirical data is therefore used to deduce the start and end of the green interval.

Till the end of this thesis, Protschky et al. did not prove the validity of the used inverse Gaussian distribution for different real-world intersections and demand scenarios. The author of this thesis assumes that especially on coordinated green waves and actuated intersections, crossing events could not

Author, Year	Exploited features, cycle length estimation	Exploited features, green interval estimation	Sampling Interval	Data amount	Primary data source	Signal strategy	Real-world validation
Kerper, 2012	Acceleration events	Crossing events and stop seconds	1 [sec]	5 %[1]	Simulation	Fixed-time	No
Krijger, 2013	Crossing events on conflicting movements	Crossing events	1 [sec]	0-50 %[1]	Simulation	Fixed-time	Yes
Schönauer, 2014	Stop seconds	Stop seconds	1 [sec]	> 100 trajectories[2]	Real-world (non-commercial)	Fixed-time	Yes
Fayazi, 2015	Acceleration events	Crossing events and stop seconds	10-80 [sec]	4289 trajectories[2]	Real-world (non-commercial)	Fixed-time	Yes
Protschky, 2015	Crossing events on conflicting movements	Crossing events	1 [sec]	40-100 trajectories[3]	Simulation	Fixed-time	No
Protschky, 2015	Crossing, stop, acceleration events		1 [sec]	30 trajectories[4]	Real-world (non-commercial)	Fixed-time	Yes

[1] Floating Car Data penetration rate
[2] number of required trajectories on a single movement to infer the cycle length and the green interval of a fixed-time controlled intersection
[3] number of required trajectories on two movements to infer the cycle length of a fixed-time controlled intersection
[4] number of required trajectories on a single movement to infer the cycle length of a fixed-time controlled intersection

Table 2.1: Overview of related research

follow strictly an inverse Gaussian distribution. Furthermore, the presented green interval estimation method [62] have not been validated by ground truth signal states which makes it difficult to evaluate the applicability of the proposed approach. Table 2.1 summarizes the used data situation of all introduced related research.

2.5 Need for Research and Methodology

With respect to the introduced literature review, further need for research has been identified that is addressed in this thesis, summarized in the following.

- **Development of a comprehensive analysis method:**

  The literature review has shown that till now research has mainly focused on the analysis of trajectories, whereas the origin, properties and methods to process this data is mainly not taken into account. Thus, this thesis contributes to the research topic of signal timing parameter estimation and signal state prediction by describing comprehensively all necessary tasks that are required to work with real-world FCD trajectories. Furthermore, all previously introduced methods make use of a modulo operation to project trajectory related events into the time scale of the cycle length, without explaining explicitly the background requirements which should therefore be investigated.

- **Signal parameter estimation with less data requirements:** The general topic of signal timing parameter estimation and state prediction based on FCD trajectories has been examined in the more recent past

by four researchers Schönauer et al. [72], Axer et al. [8], Fayazi et al. [26] and Protschky et al. [62], whereas research considered very different datasets. Figure 2.1 makes clear that no research is known that focused on a real-world, commercially available FCD source. Further research is thus needed to analyze the data capabilities of such a real FCD source. Moreover, the combination of sparse and likewise low-frequently sampled trajectories has not been analyzed in the past. A sparse trajectory volume means in the context of this thesis a single FCD source that is capable to provide for an aggregated yearly workday period only some hundreds of trajectories per hour. With other words, a sparse FCD source has in tendency a penetration or equipment rate which is much lower than 1 %. According to this, research is needed to explore a method which is capable to estimate the cycle length by considering only sparse trajectory volumes of an individual movement. In addition, the literature review allows to state that there is a need for a robust, purely data-driven approach that tries to estimate the green interval for individual movements under the mentioned data situation. Moreover, Figure 2.1 clarifies that further research should be carried out that explores the influences of traffic actuated signal control in the context of a FCD-based traffic signal state prediction.

- **Integrating intergreen times and conflicting movements:** Till now and to the very best knowledge of the author, methods for the green interval estimation did not consider conflicting movements to prove and adjust estimated green intervals against each other. Thus, the research of this thesis contributes by presenting a concept that determines firstly necessary intergreen times between conflicting movements by analyzing the road graph structure at the intersection. Secondly, this thesis presents a method that checks green interval estimates with the help of a pair-wise verification that compares green interval of conflicting movements with respect to the calculated intergreen times.

- **Transparent real-world calibration and validation:** The analysis of the related research allows moreover to state that there is a necessity for a very thorough calibration and validation in which estimated signal timing parameters are used to predict traffic signal states of real-world intersections. Thus, this thesis has the objective to present transparently achieved prediction quality with respect to the available real-world data situation by considering different statistical measures.

CHAPTER 3

Fundamentals

3.1 Characteristics of Empirical Floating Car Data

The Floating Car Data source used for this thesis has been obtained from the ADAC (General German Automobile Club), an automobile club in Germany that utilizes FCD to provide members and public users traffic related services, like level of service (LOS) and congestion estimations on German motorways [1]. As it is the main objective of this thesis to establish a data-driven signal timing estimation process that exploits commercially available FCD, the methodology needs to be capable to operate with the provided trajectory volumes, sampling intervals and other data characteristics of this data source. Therefore, this section introduces briefly an analysis that demonstrates the data situation for two intersections in Munich (Germany): Ingolstädter Straße / Frankfurter Ring and Dachauer Straße / Georg-Brauchle-Ring.

Both intersections service on the major and minor roads at least a daily traffic volume of 30 - 45 k vehicles [51]. Due to this high traffic volume, results shown in the following section represent an optimistic figure of the available FCD trajectory volumes. Roads with minor functional classes or

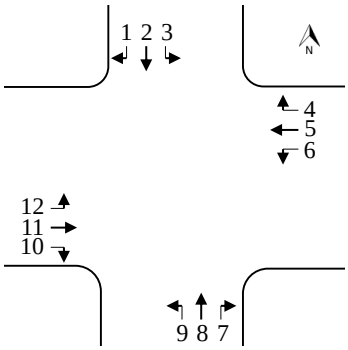
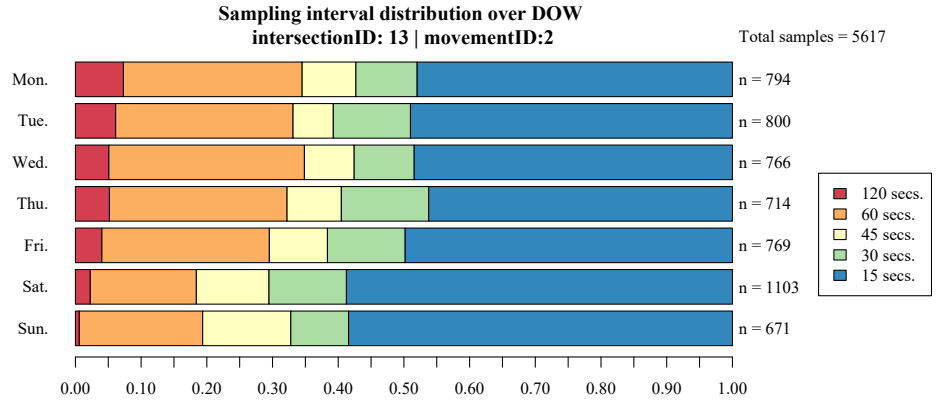
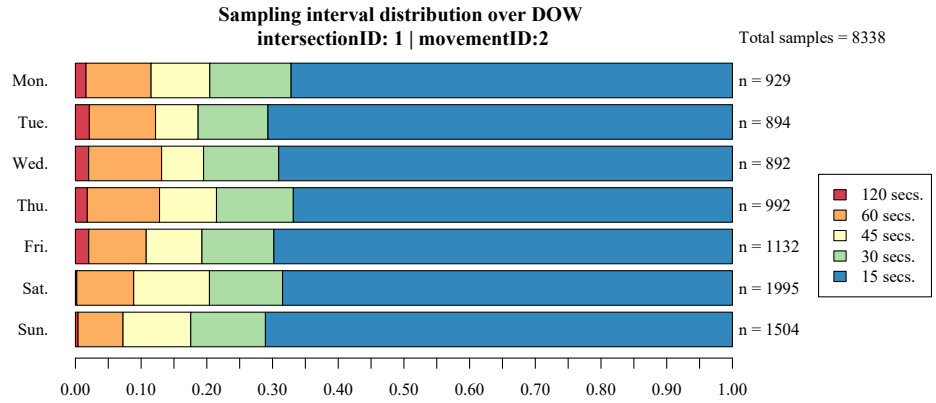


Figure 3.1: Movement scheme



(a)

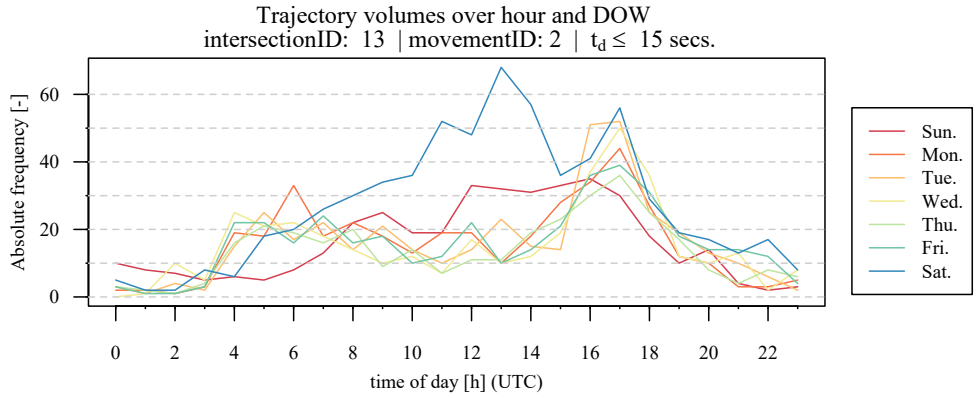


(b)

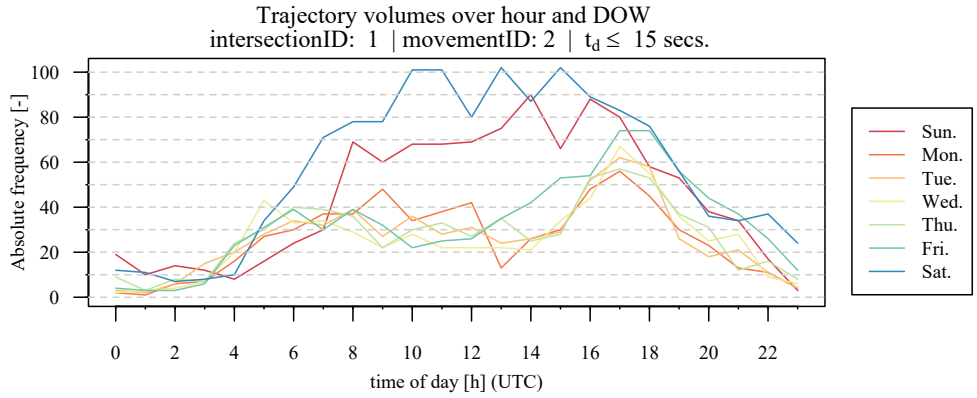
Figure 3.2: Relative shares of different sampling intervals over DOW - Ingolstädter Straße / Frankfurter Ring (a), Dachauer Straße / Georg-Brauchle-Ring (b)

turning movements will exhibit typically only a small subset of the shown trajectory volumes. In order to describe and reference results, Figure 3.1 introduces a movement numeration schema, that starts per definition in the northern approach. The data analysis covers a time period of one year (03/2014 - 03/2015).

Figure 3.2a and Figure 3.2b illustrate the distribution of collected trajectories for different days of the week (DOW) and sampling intervals t_d for the one year time period. When considering all available sampling intervals, 5617 respectively 8338 trajectories were captured on the through movements $m = 2$ at both intersection. Thus, in mean approximately 15-20 trajectories are traversing these particular movements on an individual day. Appendix B presents yearly FCD volumes of additional movements at both intersections. The analyzed data source provides five different sampling intervals whereas the sampling interval distribution differs between the tested intersections. 50-70% of all trajectories show a sampling interval of 15 secs.



(a)



(b)

Figure 3.3: Trajectory volumes - Ingolstädter Straße / Frankfurter Ring (a), Dachauer Straße / Georg-Brauchle-Ring (b), $t_d = 15$ secs

Apart from the absolute trajectory volume, the sampling interval plays a key role in the trajectory analysis, because lower sampling intervals allow a more precise motion capturing of vehicle's trajectory. To describe the trajectory continuously over time, motion is typically interpolated (e.g. linearly). A larger sampling interval lead to a larger interpolation error that causes erroneous position estimates. This phenomenon is explained in details in Subsection 4.2.1. Considering this aspect, the estimation process developed in Chapter 4 ideally requires trajectories with a sampling interval of $t_d = 15$ secs. Figures 3.3a and 3.3b illustrate therefore the available trajectory volumes for each DOW and the time of day with respect to the required sampling interval. It can be stated that the supplied trajectory volumes range from 20 - 70 trajectories per hour for each DOW, depending highly on the time of day.

3.2 General Concept and Constraints

The signal timing parameter estimation and signal state prediction approach of this thesis is well adapted to the previously introduced data conditions of a single commercial FCD provider. This adaption requires highly specialized processes in order to analyze the supplied sparse and low-frequently sampled trajectory volumes. The general concept of this processing is divided in three fundamental stages that should be briefly introduced in the following paragraphs.

The first stage covers the provisioning of a road graph model that has been obtained by the crowd-sourced map database of Open Street Map (OSM) [33]. The default XML (Extensible Markup Language) representation of the road network has been converted into a symmetrical directed graph, to enable later the implementation of a routing-based map matching. This preprocessing is realized with the help of OSM2PO (Openstreetmap Converter and Routing Engine for Java) [53], a widely used freeware software application [20, 45]. The road graph model is augmented by a manually defined interlayer that gathers single graph edges into edge sets for each movement (left, right, through). Each movement describes a sequence of inbound and outbound edges that needs to be traversed by a vehicle. The manually defined edge sets allow to query relevant trajectories for a certain movement of interest from a spatial database that stores processed trajectories. This augmentation process covers also the manually annotation of the stop bar information that is required for the signal parameter estimation. The decision to use a manually defined stop bar information is justified by the argument that the stop bar position is a very sensitive parameter for the developed estimation approach. Until now and to the very best knowledge of the author, a fully automatic and likewise reliable data-driven inference of this parameter from GPS trajectories or satellite images has not been developed respectively validated and is by itself subject to research [55, 64].

In the second stage, a map matching and path inference algorithm aligns unprocessed GPS trajectories and references the spatiotemporal movement of a vehicle to the road network.

The third stage involves the signal timing estimation process by analyzing map matched trajectories in the proximity of intersection stop bars. As trajectory volumes are typically very sparse, the estimation process requires an efficient data aggregation that exploits the principle of a cyclic traffic regulation. Folding a historical time period with respect to a particular cycle length C and a reference time point t_{ref} projects all observed trajectories to the time scale of the cycle time, illustrated in Figure 3.6. This folding process causes an increased trajectory density that allows the later estimation of signal parameters, like the cycle length and the green intervals. The approach applied assumes a signal control that is operating according to a backward calculation method [56], whereas the principle of backward calculation en-

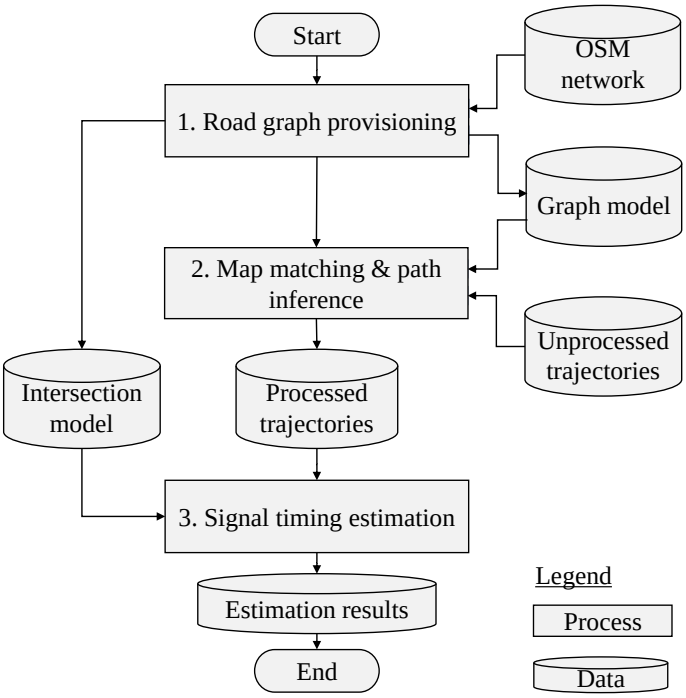


Figure 3.4: Simplified processing stages

sures that each cycle, served by the signal controller is synchronized to a standardized reference time point t_{ref}. An OCIT conform signal controller uses as a time convention the UNIX (Uniplexed Information and Computing System) time in which the timestamp "1970-01-01 00:00:00 (UTC)" is equal to zero [75]. Referring to the OCIT Outstations standard [56], a reference time point can be a globally constant value such as "01.01.1970 00:00:00 (UTC)" or a daily repeating time point like midnight "00:00:00 (UTC)". Other, widely used reference time points are specified in [56] and can also be found in Appendix F of this thesis. Thus, the signal traffic control is correctly synchronized, if the cycle second of the internal cycle counter t_{contr} is the same as the time difference since reference time t_{ref} modulo the cycle length C in charge (see Equation 3.1). If for instance a cycle length needs to be switched from C_1 to C_2 at a specific moment in time t_{switch}, this switching takes place with respect to the reference time t_{ref} [41]. Figure 3.5 illustrates this behavior in which at time t_{switch} another signal plan (Plan 2) should be used. The signal controller checks whether phases of Plan 1 and Plan 2 are matching and requirements of minimum green intervals are satisfied. If not, a transition period is used to fulfill the signal plan switch. The newly selected signal plan (Plan 2) and its cycle sequence are thus again synchronized to the specified reference time point t_{ref}. Regarding possible transition modes, used in practice, the thesis refers to the official guidelines [28, 41] which explain into details available techniques like Dwell, Max Dwell, Add, Subtract and Shortway.

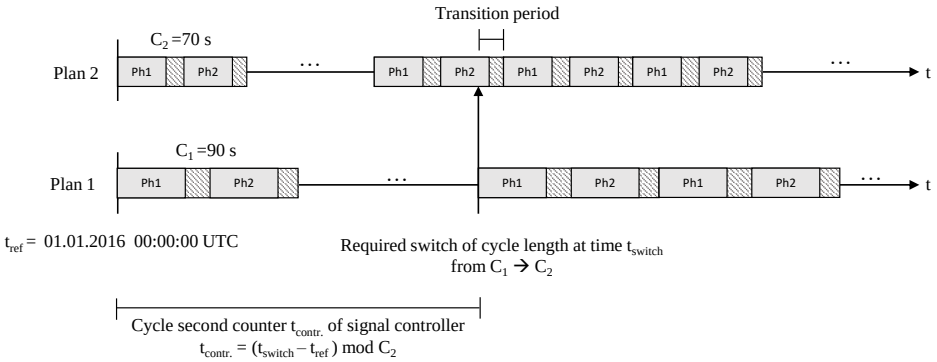


Figure 3.5: Cycle synchronization and principle of signal plan transition

$$t_{contr.} = (t_{switch} - t_{ref}) \mod C \tag{3.1}$$

The principle of time folding, illustrated in Figure 3.6, requires the knowledge of the true reference time point t_{ref} in order to fold the absolute timeline in such a manner that all historically observed cycles are stacked up congruently. Thus, if one would assume a traffic control that has a global constant reference time point like "01.01.1970 00:00:00 (UTC)" whereas signal control is in fact synchronized on a daily level at midnight "00:00:00 (UTC)", cycles would not be folded congruently, if the traffic light's synchronization takes place in a second that is not a multiple of the cycle length in charge.

The commercially obtained FCD source requires the trajectory aggregation of approximately one year in order to achieve a reasonable data density. This yearly data aggregation distinguishes between a workday and weekend subset. These two subsets are again aggregated over one hour time intervals for which typical signal parameters should be estimated. Thus, the described principles

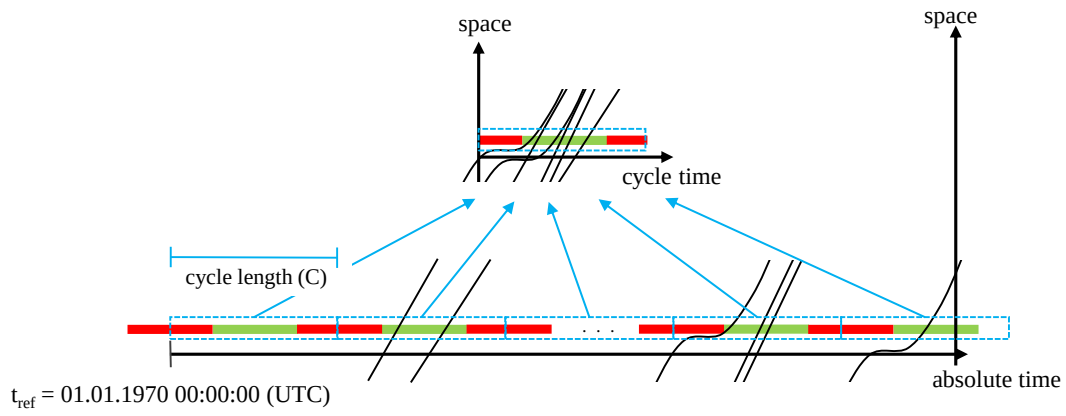


Figure 3.6: Principle of time convolution

of time folding and data aggregation imply a set of assumptions to allow the estimation of real-world signal timing parametrization.

- Unchanged signal plan set for the aggregation period (e.g. one year)
- Recurring signal plans for the time of day during workday and weekend periods
- Mainly fixed cycle length for each signal plan
- Utilization of a standard backward calculation method [56]
- Recurring traffic demand on the same daytime

Referring to the previous list, the signal timing estimation methodology of this thesis is primarily designed for fixed-time controlled intersection with a time of day or traffic demand related signal plan selection for the workday and weekend periods. Nevertheless, simulative and empirical results of this thesis prove that the inference of signal parameters can be deduced even for actuated intersections with less precision. To allow the reconstruction of typical signal parameters for a given hour in the workday or weekend period, the traffic demand needs to be similar for the same time of day. If these requirements are satisfied, the estimation methodology proposed in this thesis provides as an output three fundamental attributes to describe the signal timing parametrization that allow a linear extrapolated signal state prediction for an individual movement and a given hour of the day.

- Reference time point t_{ref}, as specified in backward calculation method
- Cycle length C, valid for a given hour of the day
- Green interval tg within the cycle time for a specific movement and a given hour of the day

3.3 Statistical Indicators

The thesis uses mainly two kinds of statistical concepts to evaluate the quality of estimated signal timing parameters. The first concept originates from Bogenberger et al. [13], who developed a methodology to evaluate the quality of real-time traffic messaging services that provide navigation systems congestion warnings on motorways. Basic idea of this concept is the superimposition of an estimated and a ground truth traffic information in which two quality indices called QKZ_1 and QKZ_2 are determined to express the overall quality of the estimated traffic information with respect to the ground truth.

The signal timing parameters and their predicted signal states are interpreted in analogy as messages of a traffic service that communicates the signal

state of a movement to a user, i.e. the driver. Bogenbergers's quality assessment distinguishes three variables, whereas variable A describes in the use case of this thesis a set of green interval estimates, E denotes a set of ground truth green intervals, derived from a simulation or from real-world traffic signal controllers. Variable D is the intersection of all predicted time intervals A and the ground truth interval E. All mentioned variables are illustrated in Figure 3.7. In analogy to Bogenberger et al. [13], quality indices QKZ_1 and QKZ_2 are calculated in this thesis as follows.

$$A = \{tg_1 \dots tg_n\}, \quad E = \{ttruth_1 \dots ttruth_n\} \tag{3.2}$$

$$tg_n = [start, end], \quad ttruth_n = [start, end] \tag{3.3}$$

$$\delta tg_n = end_n - start_n, \quad \delta ttruth_n = end_n - start_n \tag{3.4}$$

Each green interval tg_n is described as a tuple that has an interval start and end, measured in seconds. The measure QKZ_1 is therefore the detection rate that describes the proportion to which estimated green overlaps the ground truth green. The second measure QKZ_2 is synonym to the false alarm rate that describes the proportion of estimated green intervals that are not overlapping the ground truth. QKZ_1 and QKZ_2 range from zero to one (0 % to 100 %). The measures have been selected in this thesis as they allow a fast and likewise intuitive comparison of estimated and ground truth intervals for large time periods. Because both values are interconnected with each other they are typically illustrated in a two-dimensional diagram in which QKZ_2 is plotted on the abscissa and QKZ_1 on the ordinate. The best result is reached if $QKZ_1 = 1$ and $QKZ_2 = 0$.

$$QKZ_1 = \frac{\sum_{\forall i} \delta td_i}{\sum_{\forall i} \delta ttruth_i}, \quad QKZ_2 = 1 - \frac{\sum_{\forall i} \delta td_i}{\sum_{\forall i} \delta tg_i} \tag{3.5}$$

Bogenberger et al. defined in their research so called circular areas to map a value pair of QKZ_1 and QKZ_2 to a particular quality level that ranges from A to F, whereas A is the best quality level. A closer look at Bogenberg's quality level definition demonstrates that level C is covering a very large false alarm rate interval QKZ_2, ranging from 10 - 75 %. Assuming a predicted green

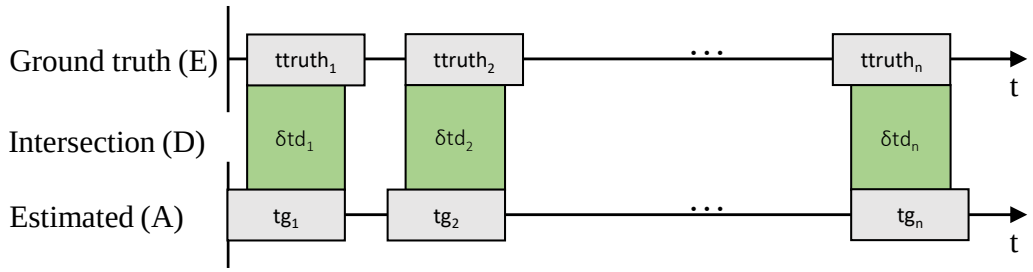


Figure 3.7: QKZ variables with respect to estimated and ground truth green intervals

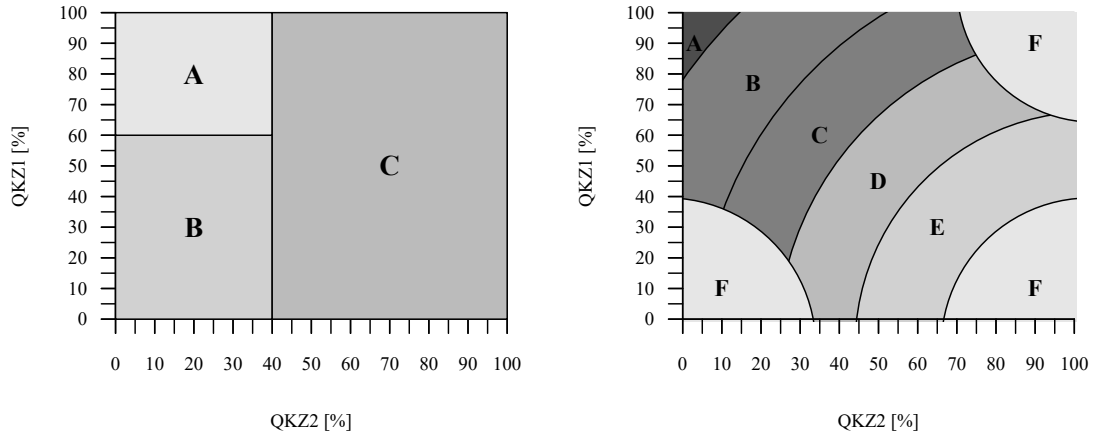


Figure 3.8: Used QKZ levels of this thesis (a) and original QKZ levels [13] (b)

interval that covers 20 secs, one would accept that 15 out of 20 predicted seconds are in reality not correct. Bogenberg et al. derived the quality level definition empirically and in coordination with different traffic message service providers that had the use case to communicate reliable congestion warnings. Because research of this thesis is evaluating QKZ parameters for predicted green intervals, an own quality level definition has been defined that reduces the number of quality levels and handles likewise QKZ_2 values more strictly. The introduced level definition divides the coordinate system only into three rectangular areas, illustrated by Figure 3.8a. According to this definition, a QKZ coordinate that exceeds the false alarm rate of 40 % is assigned to quality level C. Estimated green intervals of this level are assumed to have no essential value of benefit. In comparison to this, level A describes the most valuable green interval estimates. According to the author's opinion, even results of level B can be beneficial as predicted signal states can be used to optimize non-safety critical strategies, such as engine start-stop-systems.

Nevertheless, it should be mentioned that every aspiration level is highly related to the proposed use case of the derived information. This thesis has mainly the objective to explore the potentials of a traffic signal state prediction when using FCD independently to a specific use case. Thus, newly introduced QKZ quality levels are basically used at the later calibration and validation process to allow the readers an intuitive and easy interpretation of the result distribution.

The main disadvantage of QKZ_1 and QKZ_2 is the missing ability to display easily an absolute error when comparing ground truth start and end of green intervals against the estimated values. To present in one graphic a high density of information, the concept of box-whisker plots [49] is used. According to Murray, [54] this approach is well known in descriptive statistics and model validation. The concept allows thereby the quick evaluation of one or multiple datasets. The principle is easy. A set of data (e.g. measured distances) is

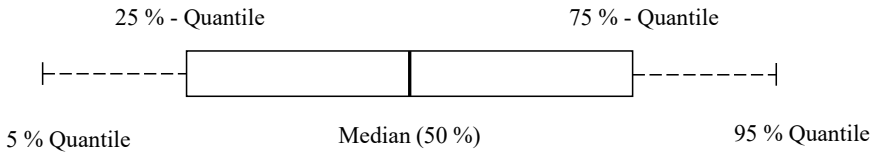


Figure 3.9: Definition of a box-and-whisker plot

ranked and divided into three different statistical quartiles Q_1=25th percentile, Q_2=50th percentile and Q_3=75th percentile. Quartile points are used to draw a three-part box, whereas the box size indicates the spread and skewness of the analyzed data distribution. Around these boxes, so called "whiskers" are used to express typically the 5th and 95th percentile of the data distribution. Measured values outside of these percentile levels are interpreted as statistical outliers [54].

CHAPTER 4

Methodology

4.1 Data Preprocessing

4.1.1 Graph and Intersection Model

Before real-world FCD trajectories can be analyzed systematically in this thesis and used to deduce signal timing parameters on intersections, a simplified model of the real-world road network is obtained to construct a graph model. This graph model is thereby used for the purpose of map matching and to decompose the trajectory flow at intersections for individual movements. A graph itself is a mathematical structure $G(V, E)$ that describes relations between objects, i.e. vertex V and edges E [11]. Both of these objects can store attributes to describe properties of the graph. As already described in Section 3.2, the thesis obtains the basic graph model from the collaborative project of Open Street Map (OSM) [33] and uses a preprocessing software that converts the default XML (Extensible Markup Language) representation into a symmetrical directed graph, to allow routing and therefore map matching operations (see Subsection 4.1.2). This preprocessing has been realized with the help of OSM2PO (Openstreetmap Converter and Routing Engine for Java) [53]. Table 4.1 presents the edge attribution used in this thesis.

All in all, the OSM road model stores already a high variety of attributes, maintained from many participants and widely used in different navigation or web services. Nevertheless, the objective of this thesis requires a very specific interlayer which is not part of the standard OSM road model. This interlayer describes for every intersection of interests a set of movements, whereas each movement stores a set of inbound (A), outbound (B) and avoiding (C) edges. Because the later signal timing estimation process needs trajectories, decomposed for individual movements, the edge sets are used to query trajectories for these individual movements. The estimation process exploits thereby only trajectories that traversed edges of the set A and B but did not traverse the avoiding edges in set C. Furthermore, the described interlayer is enriched with

Table 4.1: Used edge attribution

Attribute	Description
id	unique identification number
geographical length	edge length in unit of meter
costs	traversing costs along drawing direction
reverse costs	traversing costs reverse drawing direction
oneway	flag to describe oneway property
source	source node
target	target node
geometry	geometrical edge representation

movements' individual stop bar information, visually obtained from Google Earth [31] and stored by its coordinates, i.e. longitude and latitude. Figure 4.1 illustrates the graph model and used interlayer approach to structure intersections' movements.

4.1.2 Spatial Map Matching and Path Inference

As already described in Section 2.3 a trajectory is represented by a sequence of data points that stores information like latitude, longitude, and timestamp. Because trajectories are typically measured by GNSS devices (Global Navigation Satellite System), e.g. like GPS, the measured position can be biased by different errors [59]. Especially urban environmental conditions, such as

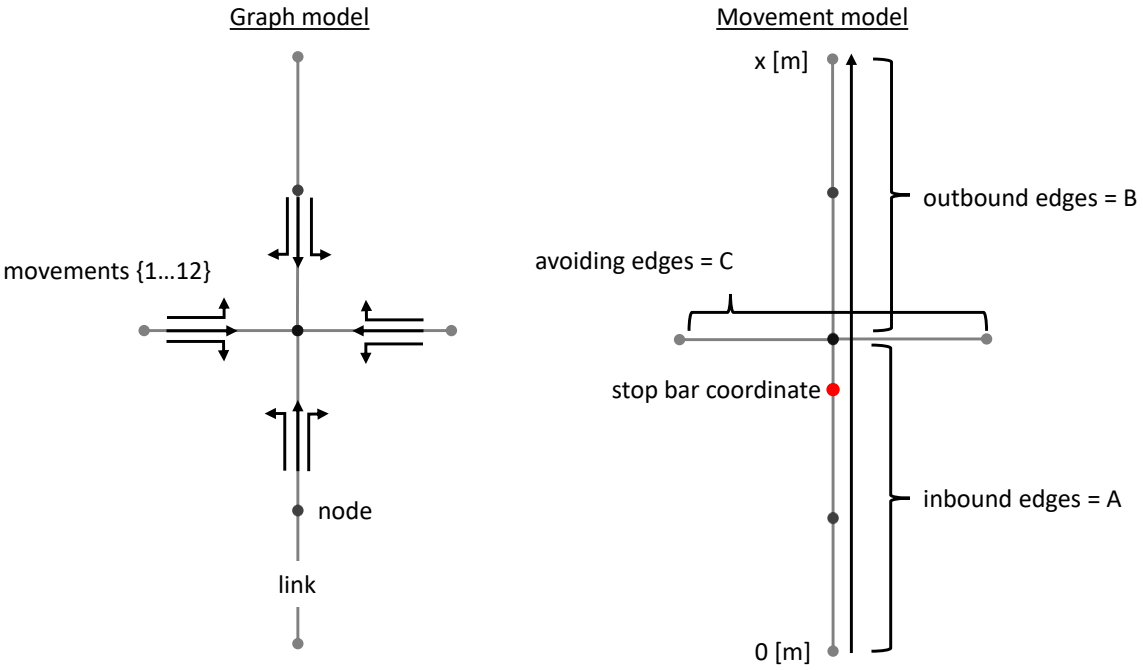


Figure 4.1: Graph and intersection model

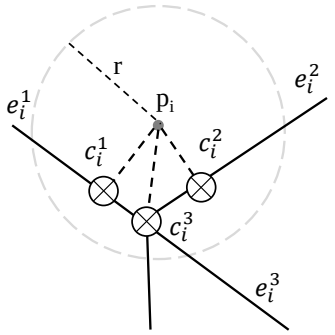


Figure 4.2: Candidate point generation

surrounding terrain, buildings and canyon walls can influence the position accuracy. Modsching et al. [52] estimate this error to be approximately 14 m in 68 % (1 σ). Depending on the data use case, a measured trajectory needs to be aligned and logically referenced with the road graph (map matching). Furthermore many applications require also an information of the most likely path a vehicle has taken through the road network (path inference) [66].

As this thesis has the objective to estimate signal timing parameters based on real-world GPS trajectories for individual movements, both problems explained need to be solved. Map matching and path inference is therefore a fundamental process in this thesis. For this important task, a slightly modified version of the spatiotemporal algorithm, described by Lou et al. [46] has been implemented. Different research has been carried out to optimize especially the runtime efficiency, well summarized in [66, 37]. However, as the basic technical principle hasn't changed essentially and the main objective of this work is not the optimization of map matching and path inference algorithm efficiency, the map matching concept of this thesis is based on the original algorithm description from Lou et al. [46]. The conceptional aspects of the algorithm and needed modifications are explained in the following.

The implemented algorithm allows the processing of trajectory data points with a sampling interval up to 5 minutes and is divided into two elementary stages. In the first stage, consecutive data points $p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_n$ of a trajectory are projected to edges e of the surrounding road network by taking the Euclidean distance x_i between p_i and e_j into account. A projected point is a candidate point c_j, where j denotes the j-th candidate of a data point p_i. This projection process is depicted in Figure 4.2. Lou et al. are assuming that the Euclidean distance between link e_j and an original data point p_i follows a zero-mean normal distribution with a sigma of 20 meters.

$$N(c_i^j) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i^j - \mu)}{2\sigma^2}} \tag{4.1}$$

$$V(c_{i-1}^s \rightarrow c_i^t) = \begin{cases} \left(\frac{d(p_{i-1} \rightarrow p_i)}{w(c_{i-1}^s \rightarrow c_i^t)}\right) & \text{if } d \leq w \\ \left(\frac{d(p_{i-1} \rightarrow p_i)}{w(c_{i-1}^s \rightarrow c_i^t)}\right)^{-1} & \text{else} \end{cases} \tag{4.2}$$

The function $N(c_i^j)$ describes thus the observation probability of a candidate point. The search radius r which is used to limit the candidate point choice set within the projection process has been set to 40 meters. Lou et al. [46] figured out that the overall map matching quality depends highly on the number of allowed candidates for one particular data point. They stated that five candidate points guaranteed a very good trade off between algorithm runtime and result quality. The same limitation has been applied in this thesis. A map matching process that would consider only the observation probability could often lead to erroneous matching results as the positional context between adjacent data points is not taken into account [46]. To overcome this problem Lou et al. introduced an other probability term, i.e. the transmission probability. The basic objective of the transmission probability is to analyze the similarity of two distance measures. The first distance measure describes the Euclidean distance between adjacent data points $d(p_{i-1} \rightarrow p_i)$. The second measure is the route length between the deduced candidates $w(c_{i-1}^s \rightarrow c_i^t)$. The route length is calculated by the help of a shortest path algorithm, such as Dijkstra [22] or A* [35]. Both distance measures are combined in Equation 4.2. As an enhancement, this thesis uses a modified version of $V(c_{i-1}^s \rightarrow c_i^t)$ that considers a reciprocal function in cases, in which the Euclidean distance exceeds the route length. This particular situation happens when candidate points are snapped to the start or end of an nearby edge, depicted in Figure 4.3.

In the second stage of the algorithm [46], a candidate graph $G'_T(V'_T, E'_T)$ is constructed for each trajectory T, whereas V'_T contains a set of candidate points and E'_T is a set of shortest path routes between adjacent candidates. This candidate graph structure is illustrated in Figure 4.4. The last objective is to search the best path P (red dashed line in Figure 4.4) from all candidate

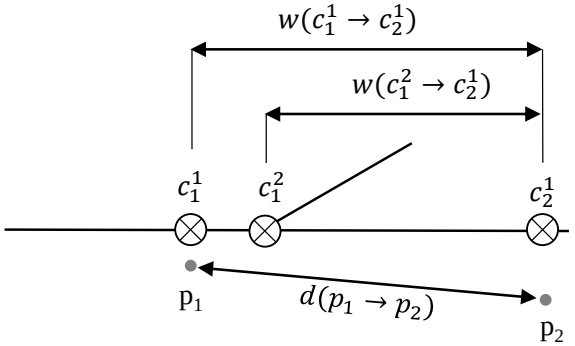
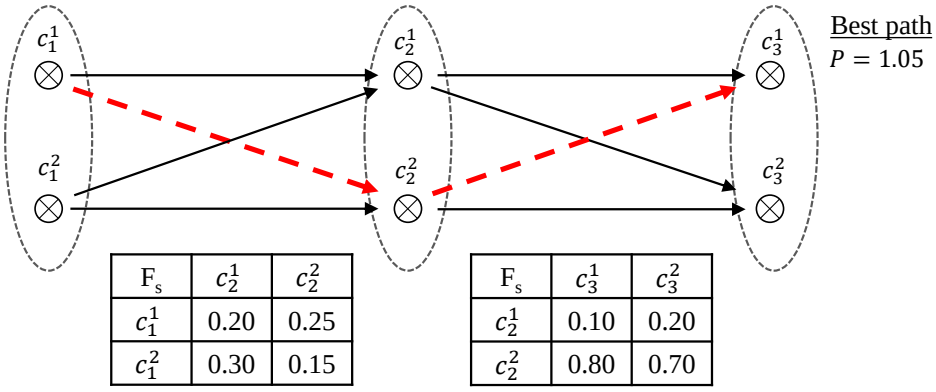


Figure 4.3: Comparison between route length w and Euclidean distance d



F_s	c_2^1	c_2^2
c_1^1	0.20	0.25
c_1^2	0.30	0.15

F_s	c_3^1	c_3^2
c_2^1	0.10	0.20
c_2^2	0.80	0.70

Figure 4.4: Directed acyclic candidate graph and its longest path

paths P_c in the candidate graph G'_T. $F(P_c)$ describes in this context the total performance of one potential candidate path P_c, whereas $F_s(c_{i-1}^s \rightarrow c_i^t)$ describes a part of the total performance sum which is generated by the transition between the source node and the target node. This process is formally described by Equation 4.5. The indices s and t are introduced to distinguish between candidate points at source s and target t nodes. This best path search problem is well known in literature under the terms of the longest path problem that can be solved for a directed acyclic graph (DAG) very efficiently with a slightly modified Dijkstra [46].

The original spatiotemporal algorithm [46] considers also a temporal probability term. This term tries to make use out of driving speed limitations on the road graph in order to distinguish in the matching process parallel roads with different speed limitations. As the network model of this thesis is only based on Open Street Map (OSM) [33], reliable speed limitations are not always available. To overcome problems of temporary missing information, the concept of temporal probability has not been implemented in this thesis.

$$F_s(c_{i-1}^s \rightarrow c_i^t) = N(c_i^t) * V(c_{i-1}^s \rightarrow c_i^t), \qquad 2 \leq i \leq n \tag{4.3}$$

$$F(P_c) = \sum_{i=2}^{n} F_s(c_{i-1}^s \rightarrow c_i^t) \tag{4.4}$$

$$P = arg\,max_{P_c} F(P_c), \quad \forall P_c \in G'_T(V'_T, E'_T) \tag{4.5}$$

The implementation of an adapted map matching and path inference algorithm leads in practice to a problematic situation which is not discussed by Lou et al. A stopped vehicle remains in reality on the same place while its GPS position starts drifting. The map matching and path inference algorithm would interpret these position changes as a driving motion in which vehicles drives forward and backward. This behavior becomes especially important in case of urban road networks, where vehicles are regularly waiting in the proximity of intersections. Depending on the measured vehicle position and the graph model, erroneous driving motions, such as loops, can induce faulty

stop bar crossing events. Due to the mentioned aspect, another preprocessing technique is needed beforehand the described map matching.

4.1.3 Processing Stopped Vehicles

To handle the explained situation, a process has the objective to identify and to cluster potential data points that belong to a stopped vehicle. For a very similar use case Tran et al. [76] enhanced the conventional and well known DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) [24] to be able to cluster stopped data points of a trajectory. Their so called TrajDBSCAN algorithm behaves similar to the classical DBSCAN that exploits the spatial proximity of nearby coordinates to build or extend a spatial cluster. A cluster consists of core points which allow the cluster to snap surrounding points, if other points are in proximity ϵ. A point becomes a core point, if a minimum number of points $minPoints$ is in its surrounding [24]. TrajDBSCAN considers likewise two parameters, space ϵ and a minimum stop time $minTime$. In comparison to the conventional DBSCAN algorithm, Tran et al. introduced the following two changes:

- A point is in the neighborhood of another one, if they are temporal linear and spatial interconnected.
- A point becomes a core point in a cluster, if the time difference between temporal cluster border points is greater or equal to $minTime$.

According to Tran et al. [76] a temporally linear and spatially interconnected sequence of clustered points $N_k = \{p_m, p_{m+1}, ...p_k, ...p_n\}$ needs to satisfy the spatiotemporal constraints of Equation 4.6 and 4.7. The core point of this

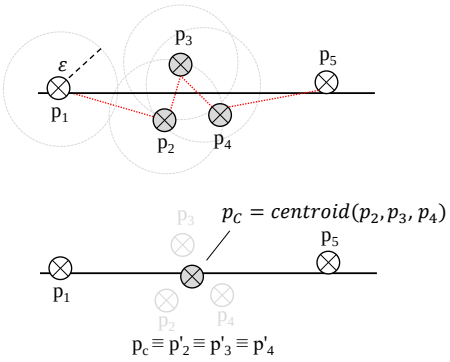


Figure 4.5: Clustering of temporal linearly and spatially interconnected data points

Table 4.2: Stored information between consecutive candidate points of a trajectory

Attribute	Description
assetid	unique identification number
source node	ID of the source node
target node	ID of the target node
source time	source timestamp
target time	target timestamp
source speed	source speed
target speed	source speed
source position	matched source coordinate
target position	matched target coordinate
edge sequence	list of traversed edges

cluster is p_k whereas t_n and t_m are defining the temporal cluster borders.

$$\forall i \quad m \leq i \leq n, t_m \leq t_k \leq t_n : distance(p_k, p_i) < \epsilon \tag{4.6}$$

$$|t_n - t_m| \geq minTime \tag{4.7}$$

Figure 4.5 illustrates the explained approach. In this example TrajDBSCAN is not able to find for p_1 a neighbor point that satisfies the spatial constraints. Thus, the algorithm tests p_2 and finds $p_3 ... p_4$ that are in ϵ proximity to each other and linearly temporal interconnected. The sequence $N_2 = \{p_2 ... p_4\}$ describes therefore a cluster, whereas clustered points may reach other neighbor points in their surrounding that allow the cluster to extend its borders. For more details, the thesis refers to Tran et al. [76]. In the use case of this work all potentially stopped data points, independent of their waiting time need to be detected and clustered, that means $minTime$ is set to zero. The dotted red lines in Figure 4.5 shows those points that are temporal linearly interconnected. Grey points indicates the found cluster structure.

Tran's et al. algorithm has been enhanced in order to merge clustered data point by a centroid calculation to omit drifting vehicle motions. That means, points $p'_2 ... p'_4$ share geometrically the same coordinates but their stored time information is preserved. The preprocessed data points are used as an input of the previously described map matching procedure. The output of this map matching and path inference generates a path P between selected candidate points. Consecutive candidate points are stored in this thesis as pairs of source and target nodes, together with all traversed edges that need to be visited to travel from the source to the target node. Table 4.2 describes the stored data structure for one source and target node pair of a trajectory.

4.2 Exploited Trajectory Information

4.2.1 Crossing Events

As vehicle trajectories are the only data source for the further estimations, this section summarizes the concept of stop bar crossing events, used for the cycle length and the green interval estimation. The applied map matching provides the knowledge of all traversed edges for each analyzed trajectory in the road network. However, an exact moment at which a vehicle has passed those edges or even the stop bar of a signalized intersection remains unknown. Therefore, the next step is to define a model, that allows to estimate vehicle motion more detailed and enables the calculation of stop bar crossing events. Every observed crossing event can be interpreted as a sensed green second in which the traffic signal that controls intersection's movement, will expected to be green. As a sampling interval of $t_d = 15$ secs is relatively long, an interpolation model is required that describes continuously vehicle's estimated position $\hat{s}(t)$ over time. In order to have a simplified data analysis and with respect to the long sampling interval, it is reasonable to assume a piecewise linear trajectory model, defined by Equation 4.9.

$$p_i = (t_i, s_i) \tag{4.8}$$

$$\hat{s}(t) = m \cdot t + b, \quad t_i \leq t \leq t_{i+1} \tag{4.9}$$

$$= \underbrace{\frac{s_{i+1} - s_i}{t_{i+1} - t_i}}_{=m} \cdot t + \underbrace{s_i - \frac{s_{i+1} - s_i}{t_{i+1} - t_i} \cdot t_i}_{=b = s_i - m \cdot t_i} \tag{4.10}$$

Each data point p_i is thus a tuple that stores the time t_i and the space s_i. Based on this linear model, an assumed moment in time $\hat{t}_{cross}$ can be calculated with Equation 4.11 in which the vehicle has crossed the stop bar. Variable s_{stop} defines the stop bar position at the movement of interest.

$$\hat{t}_{cross}(s_{stop}) = \frac{s_{stop} - b}{m} \tag{4.11}$$

Every used interpolation method is an erroneous approximation of the true trajectory and its true stop bar crossing event t_{cross}. Due to the long sampling interval, the linear interpolation causes an erroneous estimate of the crossing event $\hat{t}_{cross}$. To show the negative effect of trajectory interpolation in context of the crossing event calculation, Figure 4.6 shows two identical artificial trajectories that depict two accelerating vehicles with a constant acceleration rate till maximum allowed speed is reached. The only differences between both trajectories are the moments in which data points $p_1 \rightarrow p_2$ are sampled, the sampling interval is constant. Looking at these artificial examples it can be stated that in such situations the linear model always underestimates the true crossing event t_{cross}. The estimation error t_e reaches a maximum, when the first data point is sampled during the waiting period and the second data point

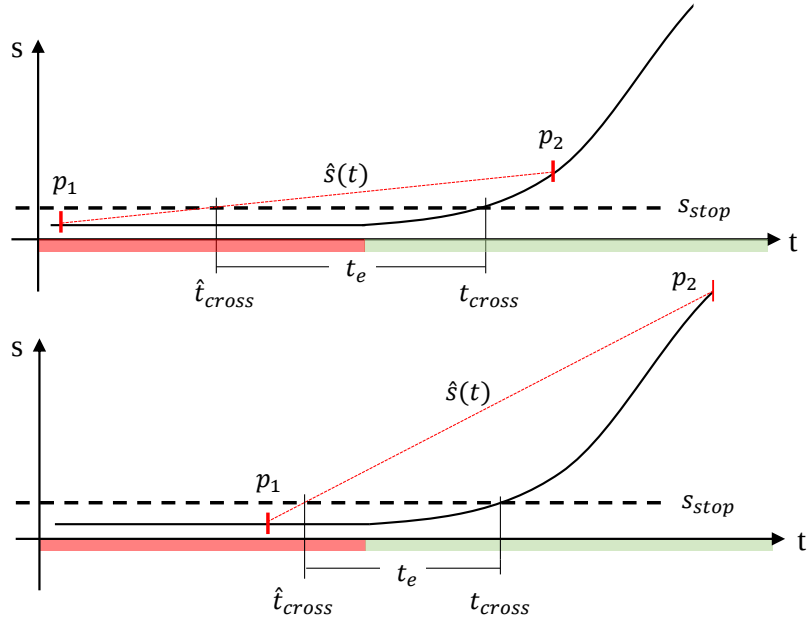


Figure 4.6: Linear trajectory interpolation

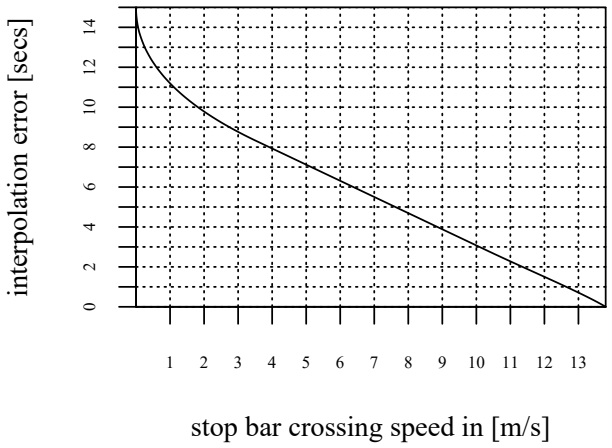


Figure 4.7: Interpolation error with respect to different crossing speeds

is generated behind the stop bar, shortly after having started the acceleration phase [7].

Referring to Axer et al. [7] lower average driving speeds during the moment of stop bar crossing have a high potential to cause large interpolation errors t_e and vice versa. To quantify this error roughly with respect to different stop bar crossing speeds, a numerical analysis has been carried out [7]. The results of this study are illustrated by Figure 4.7. It can be stated that at speeds v_{cross} of approximately 8 m/s $\approx$ 30 km/h the interpolation error t_e should be smaller

than 5 secs. Erroneous crossing events are thus a major problem in the later data processing, explained and analyzed in Section 4.3 of this thesis.

4.2.2 Stopped Vehicles

To enrich the data analysis with more information, stopped vehicles in the proximity of the stop bar can be used to sense a potential time interval in which the signal state will assumed to be red. Thus, a stopped vehicle is the complementary information of a crossing event and a stop is detected, if a vehicle is upstream the stop bar s_{stop}, has a low mean speed while being in the deceleration phase. Equation 4.12 describes formally a detected vehicle stop between two adjacent data points. Because the vehicles' position is measured by GPS, it is reasonable to observe the upstream area within a certain tolerance of $\delta s_{stop} = 35$ m. According to Equation 4.12, the in minimum measurable time period of a stopped vehicle corresponds directly to the sampling interval of $t_d = 15$ secs.

$$\delta t_{stop} = t_{i+1} - t_i \tag{4.12}$$

Stop conditions: (4.13)

$$(s_{stop} - \delta s_{stop}) \leq s_{i+1} \leq s_{stop}, \quad \delta s_{stop} = 35\ m$$

$$\bar{v}(p_i, p_{i+1}) = \frac{s_{i+1} - s_i}{t_{i+1} - t_i} \leq 5\ km/h$$

$$a(p_i, p_{i+1}) = \frac{v_{i+1} - v_i}{t_{i+1} - t_i} \leq 0\, m/s^2$$

The complete trajectory T can be allocated on a second-wise time domain to different states, as depicted in Figure 4.8. Thus, T_{state} is a trajectory state vector that distinguishes the states stopped s, driving d and the crossing c state. The time point of state s in T_{state} is denoted in this thesis with variable $\hat{t}_{stop}$ which describes an estimated stop second of a vehicle. The time point of state c denotes the previously introduced estimated crossing event $\hat{t}_{cross}$. The collection of stop seconds can be challenging as a stop second is sampled from a space interval of $\delta s_{stop} = 35$ m. In case of regular heavy tailback situations much more stopped seconds than crossing events could be collected which makes the

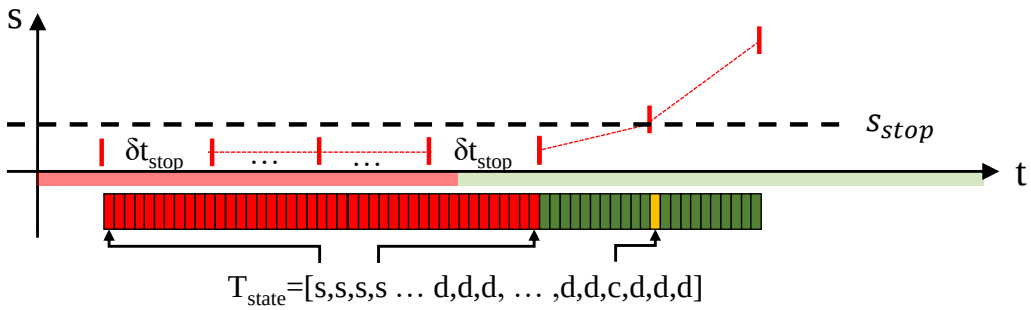


Figure 4.8: Trajectory state vector

comparability of both information difficult. Moreover, it is reasonable that a stop second nearby the stop bar is a more representative estimation of a real red second than a vehicle that is staying 20 m away from the stop bar. To make use out of this theory, each stop second $\hat{t}_{stop}$ is weighted according to its position s in the tailback. This weight is denoted in this thesis with variable w. Due to inaccurate position measurement, the position s is generalized to a multiple of 5 meters l_{slot}.

$$w(s) = \frac{1}{\lceil \frac{s}{l_{slot}} \rceil} \tag{4.14}$$

The main advantage of this approach is the self-calibrated weight that considers automatically the vehicle position. Equation 4.14 gives a formal definition of the weight w.

4.3 Filtering Erroneous Trajectories

An important aspect which has a strong influence especially on the accuracy and availability of green interval estimates, is the ability to infer and filter erroneous trajectories beforehand estimating signal timing parameters. Erroneous means in this context that map matched trajectories are causing stop bar crossing events which are not referring to a real green second. This section summarizes methods to detect and filter erroneous trajectories beforehand the cycle length and green interval estimation procedure. These erroneous trajectories can mainly be caused by four reasons, ordered with respect to their importance by the author's empirical experience.

1. **Interpolation**: To be able to calculate the moment of stop bar crossing, trajectories are modeled piecewise linearly. Lower average driving speeds during the moment of stop line crossing have a high potential to cause a faulty crossing event. A numerical analysis has already shown (Section 4.2.1) that the maximum time error (difference between true and estimated stop bar crossing) could reach the sampling interval in worst case scenarios. As aspects of interpolation are mainly related to the vehicle's acceleration phase, the author concludes that the interpolation error leads in tendency to erroneous crossing events close to the start of the green interval.

2. **Faulty positions:** Positions, measured by GPS are always affected by external influences. Modsching, et al. [52] estimate the absolute positioning error with approximately 14 m in 68 % (1σ). Additionally, the resolution of the road graph model can also influence matched vehicle position. It can be assumed that the influence of a faulty position information is not related to a certain trajectory state. Error in localization will therefore occur likewise when driving or waiting at the intersection. Thus erroneous crossing events could be probably caused everywhere in the cycle time.

Based on this assumption, the author concludes that this influence causes in tendency uniformly distributed erroneous crossing events.

3. **Different signal plans:** Within an aggregated hour h of the yearly workday or weekend period more than one particular signal plan could have been used in the past. This happens especially, if the signal switching is not taking place close to a full hour, e.g. at 09:45 h (UTC). Thus, an aggregated one hour time window can contain trajectories that belong to two different signal plans with distinct cycle lengths. The later described approach of Subsection 4.4 tries to estimate the most significant cycle length for an one hour time window of the workday aggregate. Trajectories and thus crossing events that do not belong to the inferred most significant cycle length are disturbing the cycle time nearly uniformly.

4. **Illegal crossing:** An unlikely, but also possible reason are vehicles that cross intersection's stop bar illegally during the red interval. Red light crossings could be rarely observed in everyday life, thus it was decided to neglect this cause in further analysis.

Different approaches have been tested in order to reduce the number of erroneous trajectories caused by position inaccuracy and trajectory interpolation. These approaches cover a simple trajectory classification (STC) and the utilization of machine learning algorithms like logistic regression analysis (LOGIT) and the Random Forest classification tree (RF). All methods have been benchmarked against each other at a fixed-time controlled intersection in a microsimulation environment in order to find the best method which allows to filter erroneous crossing events. Since the used simulation contains interpolation as well as position errors, is not straightforward to determine in the following analysis exactly, which erroneous crossing event has been caused by which effect. To differentiate both effects separately, two different simulation studies would have been needed. Nevertheless, at the very end of this section a visual analysis is presented that gives us the possibility to understand, which classification technique allows the best filtering of erroneous crossing events.

The decision, whether a trajectory leads to a correct or erroneous crossing event, is made on second-by-second ground truth traffic signal states, logged during each simulation second. If an estimated crossing event is not matching a true green second, it is labeled as an erroneous trajectory. The simulation study has been set up as follows:

- Utilized simulation software was AIMSUN NG whereas driving dynamics parameters refers to values from Priemer [61].

- Signal control strategy in simulation is fixed-time, with a cycle length of $C = 85$ secs and a green interval that covers cycle seconds 1 to 43. The amber interval is 2 secs and is counted as green.

- In total 9679 trajectories with a sampling interval of $t_d = 15$ secs have been simulated.
- GPS error $E_{GPS,k}$ for each trajectory k is modeled according to Equation 4.15.
- The trajectory dataset has been divided into two portions, a learning ($60\% = 5808$) and a validation ($40\% = 3871$) dataset. The validation dataset covers in total 685 erroneous trajectories (≈ 18 %) that do not match the true green interval. The learning dataset contains 1076 erroneous trajectories (≈ 18 %).

$$E_{GPS,k} \sim norm(\mu = 0 \text{ m}, \sigma = 14 \text{ m}) \tag{4.15}$$

As it is the main objective of this simulation study to deduce a classification schema that can even be used in real-world field tests, only trajectories on through movements have been taken into account. The decision to simulate and analyze only through movement traffic is based on the fact that the type of turning-movement (permissive/protected) is unknown for the complete signal timing estimation process. However, permissive and protected turning movements need to have very different driving characteristics, because vehicles could naturally stop behind the stop bar and give way to oncoming traffic. Thus, any inferred filter schema of this study should be only applied on this kind of movements. Anyway, the following subsections summarize briefly filter theory and most important findings for through movement traffic that refers to the validation data set.

4.3.1 Trajectory Classification with STC

Axer et al. [5] introduced an approach to infer erroneous trajectories on through movements of signalized intersections. The Simple Trajectory Classification (STC) approach assumes an unimpeded trajectory flow during the green interval. Therefore, vehicles should not stop behind the stop bar, otherwise stops have to be caused by an inaccurate position information. In combination with the introduced aspect of mean speed at the moment of stop bar crossing, this method will be the baseline of further analysis. According to Subsection 4.2.1 the expected interpolation error should be at $v_{cross} \geq 30$ km/h less than or equal to 5 secs. Table 4.3 summarizes the used classification schema in which only trajectories with the $states = \{1, 12\}$ are used for further data processing as they indicate a stop bar crossing speed higher than 30 km/h and no stops behind the stop bar. It is obvious that such strict filter rules can cause a very high number of rejected trajectories.

To assess the quality of each tested classification method, the concept of a confusion matrix is introduced. The confusion matrix allows thereby to visualize important performance indicators in the classification analysis [34].

Table 4.3: Simple Trajectory Classification

$v_{cross} \geq 30$km/h	Upstream stop	Downstream stop	State
TRUE	FALSE	FALSE	1
TRUE	TRUE	FALSE	12
TRUE	FALSE	TRUE	13
TRUE	TRUE	TRUE	123
FALSE	FALSE	FALSE	2
FALSE	TRUE	FALSE	22
FALSE	FALSE	TRUE	23
FALSE	TRUE	TRUE	223

Based on it, the True Positive Rate (TPR) and the False Positive Rate (FPR) can be calculated and used to conclude the classification quality. In detail, the matrix distinguishes between four different indicators. True positives (TP) are cases in which one has predicted to trust the trajectory and the ground truth confirms this estimation because the trajectory leads to a correct crossing event. True negatives (TN) are thus cases in which the classification predicted correctly the rejection of a trajectory. False positives (FP) describe cases in which one has predicted to trust a trajectory, whereas the trajectory should be better rejected. A false negative result (FN) is the rejection of a trajectory that was in fact a correct crossing event. Thus, TPR and FPR are calculated according to Equation 4.16.

$$TPR = \frac{TP}{(FN + TP)} \tag{4.16}$$

$$FPR = \frac{FP}{(TN + FP)} \tag{4.17}$$

Table 4.4 confirms the made statement that the STC schema rejects many trajectories, that are in fact correct (FN). Thus, the TPR is only about 0.47, whereas FPR is approximately 0.03. The main advantage of this simple classification approach (STC) is its ability to exclude nearly all problematic trajectories. As a negative side effect, a very high number of useful data is also rejected. These unfavorable rates are the main reasons to consider more sophisticated classifications methods.

Table 4.4: Confusion matrix: Simple trajectory classification (STC)

		Predicted	
		$0 = Reject$	$1 = Trust$
Ground Truth	$0 = Reject$	$TN = 666$	$FP = 19$
	$1 = Trust$	$FN = 1689$	$TP = 1497$

4.3.2 Trajectory Classification with LOGIT

To achieve a better classification performance, two state of the art classification methods have been examined - logistic regression (LOGIT) and Random Forest classification tree (RF). The objective of both methods was therefore to build classification models that allow, based on predictor variables $X_1 ... X_n$ linked to the ground truth response variable $Y = \{0, 1\}$, to conclude situations under which trajectories should be rejected and not used in further data processing. Coding of response variable $Y = 1$ means that an observed crossing event $\hat{t}_{cross}$ has been taken place in a moment in which the signal state was green and vice versa. As already introduced in Subsection 4.2.1, geometrical characteristics and the trajectory mean speed at the moment of stop bar crossing have an influence on the interpolation error. Fortunately, a trajectory is rich of information that can be calculated or measured before and after the crossing event to classify the observed situations. Indices up and $down$ denote in the following a measure that has been derived upstream or downstream the stop bar.

- Vehicle speed [km/h]: v_{up}, v_{down}
- Distance to stop bar [m]: d_{up}, d_{down}
- Estimated crossing speed [km/h]: v_{cross}

The concept of logistic regression [21] will be analyzed as the first alternative classification approach. As the response variable is binary, the regression model is called binary logistic regression. The regression method itself is very similar to the linear regression theory, but the main difference is the level of measurement between the variables. The term $\mathrm{P}(Y_i = 1)$ describes the probability that the response variable Y is equal to one for a certain value x_i of the predictor variable X_i.

$$\mathrm{P}(Y_i = 1) = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4.18}$$

The linear regression is typically applied on continuous variables whereas the response variable in this case is discrete, i.e. binary. Thus, if a simple linear regression model is applied, it would be able to predict for extreme x_i probabilities that are not in the value range of zero and one, which is not valid. Variable ϵ_i describes in our linear model (see Equation 4.18) the residuals between the model and x_i. To overcome this problem, Cox [21] developed a transformation process that limits the prediction values in the required range by introducing the concept of odds. The odds $odds(Y_{1/0})$ are a ratio between the occurrence probability of a certain event and its converse. The calculated odds are transformed to logits by applying the natural logarithm $\ln(odds(Y_{1/0}))$. At this transformation stage, coefficients β_n are estimated by the help of the maximum likelihood estimation (MLE). The thesis refers for more details regarding MLE to the work of Hamilton [34]. Once coefficients β_n are obtained,

the predicted logit for a particular value of x_i is transformed back into a probability $P(Y = 1|X = x_i)$ according to Equation 4.19. For the use case of a binary classification [78], a trajectory will be not rejected if $P(Y_i = 1) > 0.5$.

$$\mathrm{P}(Y = 1|X = x_i) = \mathrm{P}(Y_i = 1) \tag{4.19}$$

$$odds(Y_{1/0}) = \frac{\mathrm{P}(Y_i = 1)}{1 - \mathrm{P}(Y_i = 1)} \tag{4.20}$$

$$= \frac{P(Y_i = 1)}{P(Y_i = 0)} \tag{4.21}$$

$$logit(Y_{1/0}) = \ln(odds(Y_{1/0})) \tag{4.22}$$

$$= \ln \frac{\mathrm{P}(Y_i = 1)}{1 - \mathrm{P}(Y_i = 1)} \tag{4.23}$$

$$logit(Y_{1/0}|X_i = x_i) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \tag{4.24}$$

$$\mathrm{P}(Y = 1|X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)} \tag{4.25}$$

Table 4.5 summarizes the estimated coefficients of the applied logistic regression model as $ln(odds)$ values. The analysis has been carried out with the help of the R-Statistics software [65]. Most important results can be interpreted as follows. The column $Pr(> |z|)$ represents the p-value, i.e. the so called observed significance [50], for rejecting the null hypothesis that the estimated coefficient is equal to zero. In case of this study, it can be stated that nearly all supplied predictor variables are at least significant for $\alpha = 0.05$.

Table 4.5: LOGIT coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.802065	0.326059	2.460	0.0139 *
d_{up}	-0.024967	0.007813	-3.196	0.0014 **
d_{down}	0.009914	0.007920	1.252	0.2107
v_{up}	0.073430	0.003427	21.427	< 2e-16 ***
v_{down}	-0.039717	0.006268	-6.337	2.35e-10 ***
v_{cross}	0.081838	0.032487	2.519	0.0118 *

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Only the downstream distance between the vehicle and the stop bar d_{down} has no significant influence on the model. In the opposite, the upstream distance plays a very important role in this model. The negative coefficient of variable d_{up} documents that odds of a correctly sensed green second is reduced for every additional meter between the vehicle and stop bar in mean by the factor of $exp(-0.025) = 0.98$. The very same negative relation occurs for the variable v_{down}. In contrast thereto, a higher vehicle speed before the crossing event is by far the most significant parameter in which odds for a correctly

Table 4.6: Confusion matrix: LOGIT classification

		Predicted	
		$0 = Reject$	$1 = Trust$
Ground Truth	$0 = Reject$	$TN = 457$	$FP = 228$
	$1 = Trust$	$FN = 112$	$TP = 3074$

observed green second are increased in mean by every additional km/h by the factor $exp(0.073) = 1.07$. The analysis also shows that the crossing speed is likewise an important parameter that is compared to other variables less significant.

The used set of predictor variables lead to a very good classification result, summarized in Table 4.6. The TPR of the used LOGIT model is about 0.96 which is a quite good value because it allows to preserve a high number of trajectories. As a negative side effect, the FPR is also increased to approximately 0.33. However, this model allows to reject 67 % of all erroneous trajectories. Next, the RF algorithm is evaluated to assess the filter abilities of this innovative classification approach.

4.3.3 Trajectory Classification with RF

According to Breiman [15], the developer of the RF algorithm, the main advantage of this classification technique is its unexcelled classification accuracy. However, the main disadvantage of this promising method is its black box approach [30]. Nevertheless, most important fundamentals should be introduced briefly. The RF algorithm creates multiple classification trees with the objective to generate a model that is capable to predict the value of response variable based on multiple different predictor variables [16]. Each single tree is derived on a bootstrapped sample of the provided input data. The classification output for a certain set of predictor variables is a majority vote of all generated trees. As the algorithm per design uses bootstrapping, no additional cross-validation technique [18] is needed to ensure model validity. For more details regarding this algorithm, this thesis refers to the work of Breiman [15]. The analysis has been carried by analyzing the mentioned set of predictor variables with help of the R-Statistics software [65, 44]. The following classification results used the standard software settings for the RF algorithm [44].

Table 4.7: Confusion matrix: Random Forest tree classification

		Predicted	
		$0 = Reject$	$1 = Trust$
Ground Truth	$0 = Reject$	$TN = 578$	$FP = 107$
	$1 = Trust$	$FN = 88$	$TP = 3098$

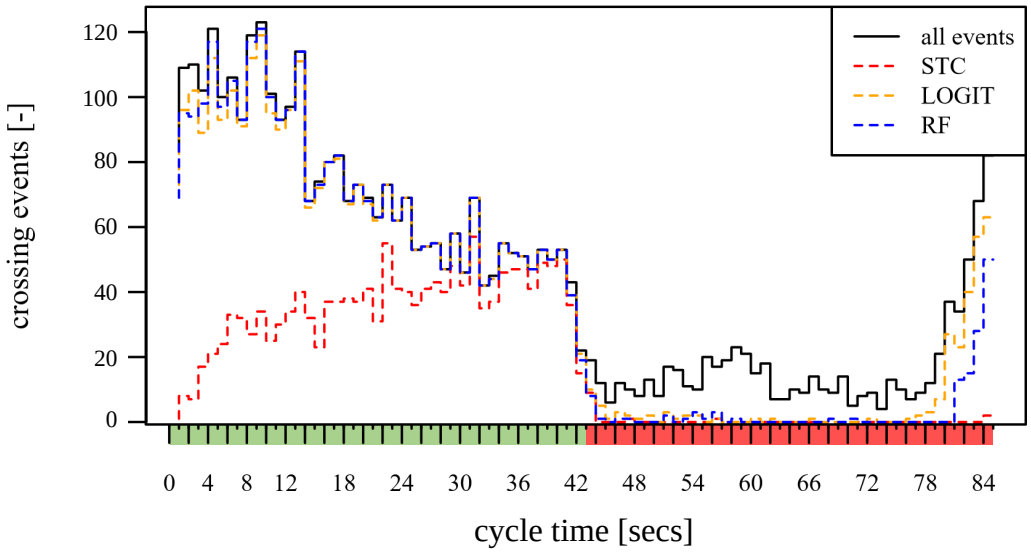


Figure 4.9: Crossing event distributions for different filtering methods

The gathered results of Table 4.7 prove Breiman's statement. The RF algorithm shows in comparison to all other methods the overall best confusion matrix in which the TPR is approximately 0.97 and even the FPR lies on a very low level of approximately 0.16.

4.3.4 Conclusion

In comparison to the tested LOGIT approach, the RF algorithm allows a reduction of the FPR by factor two. Furthermore, 84 % of all erroneous trajectories can be detected and correctly filtered. The total number of reduced erroneous trajectories is very promising, but till now the study has not present a rough conclusion, how the classification methods handle explicitly the introduced interpolation error. For this reason, Figure 4.9 illustrates the distribution of crossing events over the complete cycle time $C = 85$ secs with respect to all tested classification methods. The real green interval is indicated on the abscissa and starts at the first cycle second, whereas the solid black line in Figure 4.9 represents all counted crossing events per cycle second without having applied any filtering method. The figure shows that the solid black line starts already rising at second 78 of the cycle time.

Thus, it can be stated that the interpolation error in this simulation experiment is approximately about 8 secs. As already explained in the beginning of this section, errors caused by the position inaccuracy superimpose the complete cycle time, which can be even proved in Figure 4.9. Hence, the last 8 secs of the cycle time contain both error types. Figure 4.9 illustrates moreover the filtering impacts of all tested methods. The STC approach allows in comparison to the other methods the best reduction of the interpolation error. As a negative side effect, a very high number of correct crossing events are falsely rejected.

Because the tested RF algorithm is clearly superior to all tested classification methods, it will be selected as the best filter approach for the ongoing data processing in this thesis.

4.4 Estimating the Cycle Length

After having introduced the data input and the methods to filter erroneous trajectories, this section explains an approach to infer the cycle length of an intersection. The proposed cycle length estimation methodology has in contrast to other existing approaches (see Section 2.4) the following advantages:

1. **Movement-wise estimation:** The estimation methodologies developed by Protschky et al. [62] and Krijger et al. [43] are analyzing stop bar crossing events on at least two conflicting through movements in order to estimate the cycle length. Both approaches assume that the crossing events on these movements do not overlap each other. The main drawback of this idea is the requirement to obtain enough trajectories on multiple conflicting movements. If the number of trajectories on one movement is insufficient, this technique could easily fail. Thus, the proposed cycle length estimation methodology in this thesis has the capability to infer the cycle length information without requiring data on multiple movements at a signalized intersection.

2. **Exploiting one statistical measure:** Protschky et al. presented in [63] an other approach that is in principle capable to infer the cycle length on individual movements. Therefore, a combination of three statistical measures is used that range from trajectory correlation over variance to image analysis. The proposed method has strengths and weaknesses. On the one hand, the combination of measures allows for sure a very reliable cycle length inference. On the other hand, if a process always requires multiple input parameters in parallel to infer the cycle length, the estimation could fail or get worse, if some parameters are missing. Hence, the method presented in this thesis tries reliably to infer the cycle length only with the help of one statistical measure.

3. **More efficient data exploitation:** The research of Kerper et al. [40] and Fayazi et al. [26] is exploiting trajectory related events to calculate a single statistical parameter for an individual movement in order to estimate the intersection's cycle length. Their main weak spot is the extraction of a small trajectory subset that experienced the sequence of deceleration, waiting and acceleration in order deduce the green start for consecutive cycles. Even Schönauer et al. [72] analyzed in his methodology only a subset of trajectories that was forced to stop. Such a trajectory state specific filtering can be problematic, if the intersection is part of a green wave, in which only a small number of trajectories needs to stop. Thus, the cycle length

estimation of this thesis exploits efficiently all crossing events without focusing on trajectories that experienced very specific state sequences.

The introduced intersection model (see Subsection 4.1.1) allows the selection of trajectories for a certain intersection and its movements m according to the numeration scheme (see Subsection 3.1). Furthermore, on all through movements $m = \{2, 5, 8, 11\}$ the developed trajectory classification process is applied and allows efficiently to filter erroneous trajectories. As intersections' movements are forced to the same cycle length, its inference should be carried out on the movement $m_{best,h}$ that indicates the highest trajectory volume vol for a given hour $h = \{0...23\}$. This step ensures that the cycle length inference is always concluded from the best available data situation.

$$m_{best,h} = \forall m, h \; arg\,max_m(vol_{m,h}) \tag{4.26}$$

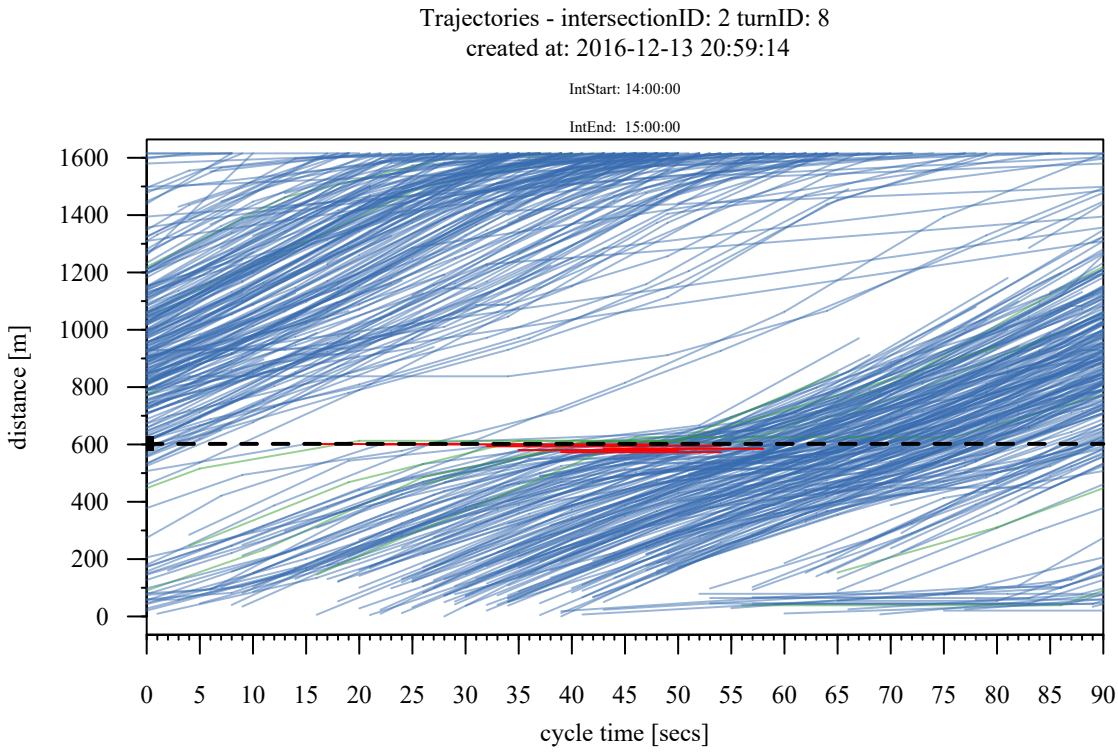
$$\hat{t}_{cross,C_p,k,m_{best}} = (\hat{t}_{cross,k,m_{best}} - t_{ref}) \mod C_p \tag{4.27}$$

Next, the principle of time folding (see Section 3.2) is applied. Thus, each trajectory T_k and therefore each crossing event $\hat{t}_{cross,k}$ is projected under consideration of a reference time t_{ref} into the time scale of a potential cycle length Cp. This projection forces all trajectories to take place in a very short time period, i.e. the cycle length. If the correct cycle length is applied, stop bar crossing events $\hat{t}_{cross,C_p}$ appear only during a small time window, i.e. the green interval. The effect of trajectory and crossing event concentration provides the fundamental basis for the cycle length estimation and is depicted in Figure 4.10a. The figure shows a real-world example of convoluted trajectories for a cycle length of 90 secs at the intersection Ingolstädter Straße / Neuherbergstraße in Munich (Germany). The dashed horizontal line indicates the stop bar. Red labeled trajectories are stopped vehicles, blue trajectories are usable trajectories that satisfy the Random Forest Tree classification process, whereas green trajectories are labeled to be erroneous and thus excluded in further data analysis. In the case of a wrong cycle length, trajectories and crossing events will occur almost uniformly distributed over the cycle with the uniform distribution parameters $(a = 0, b = C_p)$. This situation is illustrated in Figure 4.10b.

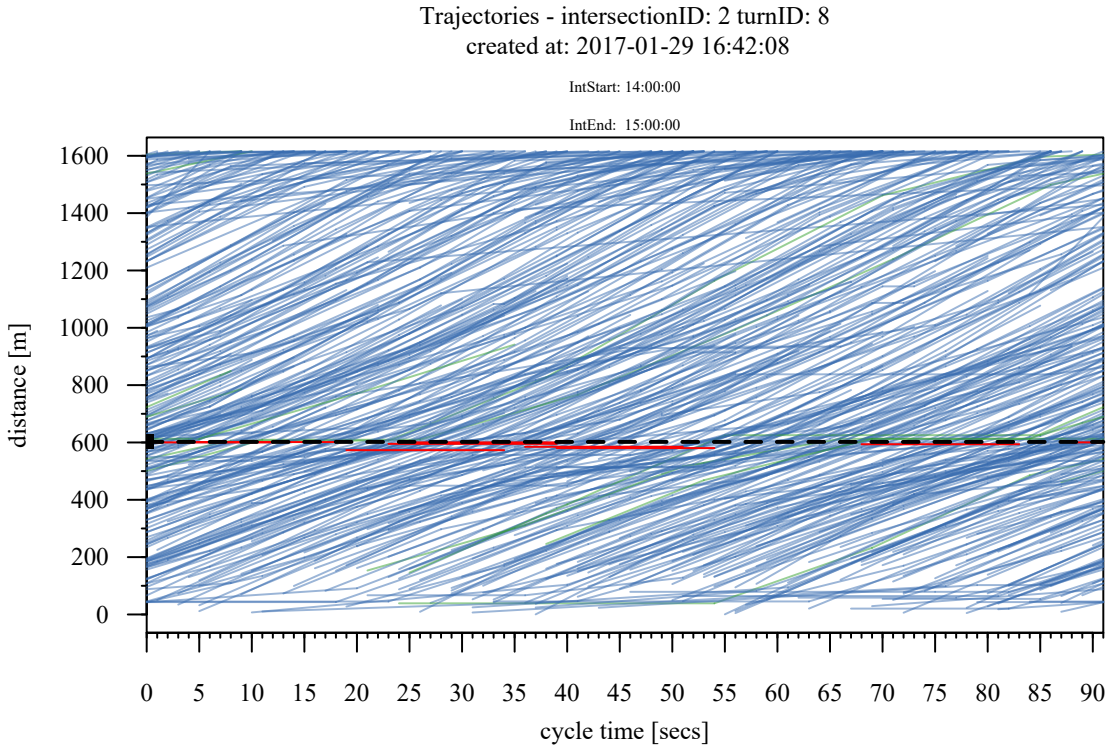
To identify the correct cycle length, the method of this thesis proposes a hypothesis test that has the objective to find a cycle length candidate C_p that causes a strong crossing event concentration. A strong enough concentration may allow to conclude that the empirical crossing event distribution deviates or differs statistically significant from the assumed theoretical uniform distribution $F_{unif}(a = 0, b = C_p)$. The tested cycle length candidates are typically in a range of $C_p = \{30...120\}$ secs that corresponds to the min/max values of German guidelines [28]. The null and alternative hypothesis are defined as follows.

$$H_0 : \hat{F}_n(\hat{t}_{cross,C_p}) \textit{ follows a uniform distribution.} \tag{4.28}$$

$$H_1 : \hat{F}_n(\hat{t}_{cross,C_p}) \textit{ follows not a uniform distribution.} \tag{4.29}$$



(a) Correct cycle length, $C = 90$ secs



(b) Wrong cycle length, $C = 91$ secs

Figure 4.10: Real-world example of time convolution

The magnitude of deviation between the empirical cumulative crossing event distribution $\hat{F}_n(\hat{t}_{cross,C_p})$ and the assumed theoretical cumulative uniform distribution $F_{unif}(a = 0, b = C_p)$ can be measured and evaluated with respect to different test statistics. This problem is also known under the term of goodness-of-fit statistics [74]. For this purpose, different measures have been developed in the past [3, 48, 2]. The latest and most advanced one is the Anderson-Darling-Test (AD-Test) [3]. The test statistic A^2 [3] is defined by Equation 4.30 in which Y_i are the sorted empirical values and $F(Y_i)$ describes the cumulative theoretical distribution function.

$$A^2 = -n - S \tag{4.30}$$

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln(F(Y_i)) + \ln\left(1 - F(Y_{n+1-i})\right)\right]. \tag{4.31}$$

The magnitude of the derived test statistic A^2 as well as the drawn sample size are used to estimate a so called p-value. Mendenhall et al. [50] define this p-value as follows:

> The p-value or observed significance level of a statistical test is the smallest value of α for which H_0 can be rejected. It is the actual risk of committing a type I error, if H_0 is rejected based on the observed value of the test statistic. The p-value measures the strength of the evidence against H_0.

The calculation of p-values with respect to the test statistic A^2 and the drawn sample size is typically carried out with methods, like the Monte-Carl-Simulation. This thesis refers to the R-Statistics software [65, 25] that includes the AD-Test and the p-value calculation based on the implementation from Marsaglia et al. [47].

According to Mendenhall et al. [50], the null hypothesis H_0 can be rejected, if the p-value is less than or equal to a preassigned significance level α. A very small p-value can thus be interpreted as a strong evidence that H_0 is false and should be rejected [50]. This thesis specifies a significance level of $\alpha = 0.005$ that is required in order to reject the null hypothesis H_0. When working with hypothesis theory, one should remember that two potential error types could occur and thus lead to a false decision (type I and type II errors).

The already mentioned type I error is the incorrect rejection of the null hypothesis [57]. A simple example is an innocent person, who was convicted by court to be guilty, whereas the null hypothesis assumes that the person is innocent. Assigned to the cycle length estimation, one would erroneously infer a cycle length based on assumed non-uniformly distributed crossing events, which is in fact not correct because the crossing events are in reality uniformly distributed. A type II error is the incorrect rejection of the alternative hypothesis [57]. This means, a cycle length that is in reality present at an intersection will not be detected by the hypothesis test, because the hypothesis

test has not found enough evidence to reject the null hypothesis H_0. This error probability of type II errors is typically denoted by the variable β [57]. As in practice α is protecting the decision process for type I errors (i.e. wrong cycle length identification), the behavior and the distribution of type II error are often neglected. However, it is necessary to know till which sample size vol_{crit} type II errors occur for the mentioned significance level of $\alpha = 0.005$. If vol_{crit} is exceeded and the hypothesis test is not able to reject the null hypothesis, the intersection of interest is properly violating the introduced requirements of signal control strategy, defined in Section 3.2 and no cycle length can be inferred with the specified significance level of $\alpha = 0.005$.

The sample size vol_{crit} is used in this thesis as the minimum trajectory volume, required to infer surely the occurrence of a cycle length at a given hour. Figure 4.12 illustrates therefore the result of an own Monte-Carl-Simulation in which the occurrence of type II errors have been simulated with respect to multiple sample sizes of stop bar crossing events and different ratios of green interval g to the cycle length C. For this analysis 200 replications i have been calculated whereas in each iteration the hypothesis test has been supplied with synthetically generated crossing events ranging from 5 to 100 samples, with a step size of 5 ($j = 100/5 = 20$). That means $i \cdot \ j \ = 4.000$ experiments have been simulated in total. A synthetic number of crossing events t_{cross} were calculated according to Equation 4.34, the ground truth cycle length C was set to 85 secs. The objective was to find the true cycle length in a range from 30 to 120 secs, if the introduced hypothesis test is supplied with different samples sizes of t_{cross}. Figure 4.11 shows the result of one particular iteration of the Monte-Carlo-Simulation in which the hypothesis test has been supplied with 90 crossing events. The horizontal red line is indicating the required significance level to reject the null hypothesis. It can be stated that at a cycle

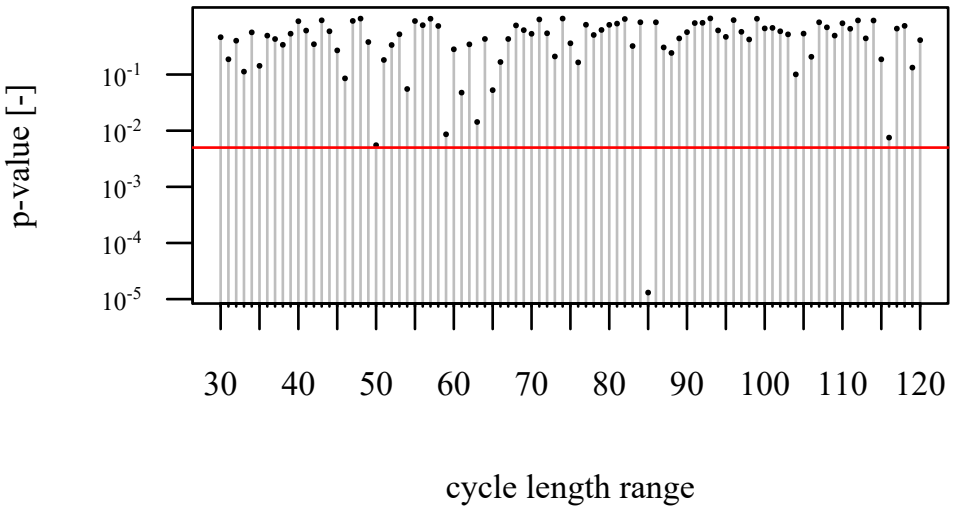


Figure 4.11: Single iteration of a Monte-Carlo-Simulation and the resulting p-value over different cycle lengths, sample size = 90 crossing events

length of 85 secs, the resulting p-value reaches a global minimum, much lower than all other cycle length candidates.

$$r \sim unif(0, g, n = 100\,k) \tag{4.32}$$

$$N \sim unif(0, 200\,k, n = 100\,k) \tag{4.33}$$

$$t_{cross} \sim r + (N \cdot C) \tag{4.34}$$

Figure 4.12 allows to state that higher ratios of green interval g to cycle length C require in general a larger volume of crossing events to reach the same probability of type II errors. This phenomenon can be reasoned, because if in 75 % of a cycle, which was in fact the true cycle, uniformly distributed crossing events can be observed, the AD-test requires a high number of samples to surely reject the null hypothesis. As only in 25 % of the tested cycle no crossing events can be observed, the crossing event distribution $\hat{F}_n(t_{cross,C_p})$ will look very similar to the theoretical uniform distribution $F_{unif}(a = 0, b = C_p)$ that needs to be rejected. At a volume of approximately 90 - 100 crossing events, type II error probability remains under the level of 0.001. As in real-world situations other external influences can disturb the crossing event distribution and the critical volume $vol_{crit} = 90$ is already a relatively small sample size to infer further information from it, this critical volume is used in this thesis like in other works [6] as the critical trajectory volume vol_{crit}.

As introduced in this section, the cycle length estimation is applied over time intervals of typically one hour $h = \{0...23\}$. Thus, it could happen that two or more cycle length candidates have been found that match the defined significance level of $\alpha = 0.005$. This situation occurs, if the signal plan switching does not take place close to a full hour (e.g. at 18:30 h). In such situations

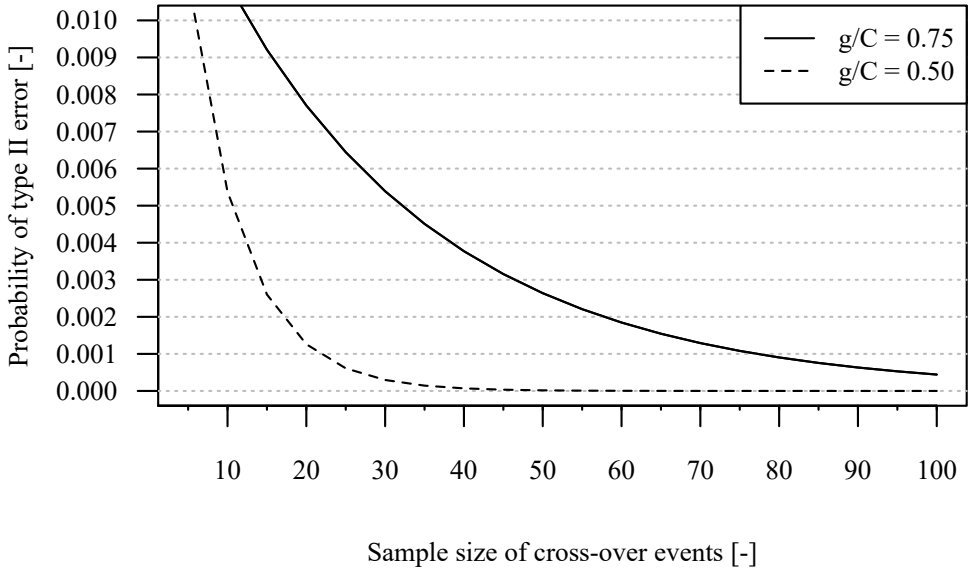


Figure 4.12: Probability of type II errors over sample size and g/C ratios

the developed methodology proposes to minimize the resulting p-value of the AD-test to find the most significant cycle length C_{best} for the tested hour. The result of this minimization is used for all other movements m.

$$C_{best} = arg\,min_{C_p}(p_{value}(A^2_{C_p})), \quad C_p = [30 \dots 120] \tag{4.35}$$

Previous explanations are assuming that the reference time t_{ref} in modulo operation of Equation 4.27 is known. In fact, the true reference time which is used for intersections' backward calculation (see Section 3.2) is unknown and needs to be determined. Fortunately, the number of potential reference time points is at least in Germany limited and defined by the OCIT Outstations standard [56]. For the United States of America, a signal traffic controller uses typically "00:00:00 (midnight)" for the synchronization [41, 17], which is conform to the backward calculation method "RRV midnight" in OCIT Outstation standard, described in Appendix F.

As already explained and illustrated in the fundamentals of this thesis (see Section 3.2), the application of a wrongly selected reference time point could cause a wrong convolution of cycles, if the traffic light's synchronization takes place in a second that is not a multiple of the cycle length in charge. Hence, if it is not possible to infer any cycle length with the described methodology at an intersection of interest, the thesis proposes to enumerate over all common reference time points according to the OCIT definition [56]. If after this enumeration no cycle length is found, the intersection might not be controlled by a fixed cycle length and thus the estimation methodology would fail.

4.5 Estimating the Green Interval

4.5.1 Simple Green Interval Estimation

The successfully deduced cycle length information and thereby projected trajectories are the base for estimating the green interval. The proposed method shows in contrast to alternative approaches, introduced in Section 2.4, two main advantages which should be explained beforehand explaining the details. Subsequent paragraphs refers mainly to the research from Protschky, et al. [62] and Schönauer et al. [72] as their works focused likewise on the analysis of sparse trajectory volumes. In contrast to the research of this thesis, both authors developed methods that are specialized to exploit high-frequently sampled trajectories with a sampling interval of $t_d = 1$ sec.

1. **Unknown event distributions:** Protschky, et al. [62] assumes, that crossing events $\hat{t}_{cross,C_{best}}$ always follow an inverse Gaussian distribution. He concluded that different intersection topologies and the occurrence of tailbacks can probably lead to other distribution shapes that should be modeled by changing the distribution parameters. Axer et al. [6] has simulated

event distributions for crossing vehicles at fixed-time controlled intersections considering two different movement's degrees of saturation. The study [6] proved that movements' degree of saturation influences intensively the shape of the crossing event distribution. Moreover, the results suggest that empirically collected crossing events can not always be modeled with an inverse Gaussian distribution. The author of this thesis assumes that Protschky's assumption holds mainly for isolated intersections at which vehicle arrivals are more or less randomly distributed. Especially on arterial roads with short segment length between consecutive intersections, the crossing event distribution can be affected by the presence of green waves and previous intersections. In such situations vehicles are passing an intersection in large platoons and the crossing distribution of vehicles could be very similar to a uniform distribution.

Based on this argumentation, the presented approach in this work does not assume any specific distribution model in order to extract the green interval from the collected crossing events $\hat{t}_{cross,C_{best}}$.

2. **Self-verification of data coherence:** As introduced in Section 2.4, Schönauer et al. [72] are exploiting second-by-second bicycle trajectories in the proximity of stop bars. Their approach focuses on the distribution of stopped cyclists per cycle second, whereas a low sampling interval of $t_d = 1$ sec allows to measure precisely the waiting time of every cyclist. Stopped cyclists per cycle second are counted in a histogram analysis in order to infer the green interval. In comparison to Protschky et al., the method is free of any distribution assumptions and the histogram analysis performed basically well. To detect the green start within the cycle time Schönauer et al. are searching the cycle second in the histogram that shows the highest drop of stopped cyclists. A closer look at their research shows that the used measure seems to be unreliable in cases, in which the total number of measured stops is very sparse. The authors did not discuss in [72] the robustness of their measure when analyzing only small number of events.

   The green interval estimation of this thesis utilizes similar to other approaches [40, 72, 62] a histogram analysis. As an improvement to previous works, this thesis develops a concept that tries to verify the data coherence of sampled trajectories. Self-verification of data coherence means in this context that the estimation process is able to operate on small numbers of crossing events until predefined constraints are violated. This violation indicates in the estimation process that data is too sparse or inconsistent. Data coherence ensures therefore that the collected events are frequently enough to form a contiguous time interval with a minimum time duration. If the data source is not able to provide enough samples to ensure these constraints, the process stops automatically and no green interval estimation is possible.

At this processing stage, the green interval estimation methodology exploits the trajectory state vector $T_{state,C_{best},k}$ of the k-th trajectory that have been projected to the inferred cycle length C_{best}. The trajectory state vector has been introduced and explained in Section 4.2.2. Furthermore, for each through movement $m = \{2, 5, 8, 11\}$, the developed trajectory classification process is applied to filter erroneous trajectories. As described in the previous Section 4.5.1, the following process is supplied in minimum by $vol_{crit} = 90$ trajectories.

In the very first processing step ❶ of the green interval estimation, the crossing events are extracted from trajectory state vectors $T_{state,C_{best},k}$ and analyzed in a histogram h_{cross}, similar to other approaches [40, 72, 62]. Each bin b of this histogram has a granularity δt of one second, which is argued on the fact that the signal state switching takes place on the same time resolution [56]. Thus a bin $b \in \{1 \dots C\}$ counts the number of observed crossing events h_{cross} for a particular cycle second.

$$c_k(b) = \begin{cases} 1, & \text{if} \quad T_{state,C_{best},k}(b) = crossing \\ 0, & \text{else} \end{cases} \tag{4.36}$$

$$h_{cross}(b) = \sum_{\forall k} c_k(b) \tag{4.37}$$

$$\delta h_{cross}(b) = \frac{h_{cross}(b) - h_{cross}(b-1)}{\delta t} \tag{4.38}$$

Such a histogram of crossing events h_{cross} is synonym to a crossing event distribution and depicted in Figure 4.13 ⓐ. Because the filtering of erroneous trajectories will never reach perfection and other external error influences may occur, it is very unlikely to achieve a crossing event distribution that allows to divide the complete cycle time into two contiguous time intervals - a green interval that contains crossing events and a red interval that contains absolutely no crossing events. Thus, it is necessary to classify each cycle second by means of the counted crossing events in order to obtain these two contiguous time intervals. Each cycle second is therefore labeled with a binary $state = \{0, 1\}$ in which state coding zero means, the cycle second belongs to the red interval and vice versa.

Before explaining more details of this classification procedure, the thesis discusses briefly an alternative approach and the main reason, why it has not been applied in this classification context. Schönauer et al. [72] followed the idea to calculate a difference quotient $\delta h_{cross}(b)$ between frequencies of consecutive bins to infer a meaningful turning point in the crossing event distribution that is used to detect the start of the green interval. When looking at Figure 4.13 ⓐ it is imaginable that sparse data would cause a very volatile difference quotient and thus a robust turning point detection would be problematic.

Thus, this thesis follows an iterative rule-based approach, illustrated in Figure 4.13. This concept uses three fundamental assumptions which were already mentioned implicitly and are itemized explicitly in the following.

1. **The red interval contains the lowest number of crossing events:** The red interval should cover with respect to all other cycle seconds, the lowest number of crossing events. Furthermore, erroneous crossing events that are superimposing the red interval are assumed therefore to be uniformly distributed and the mean frequency of occurrence should be relatively constant in this time interval. If a method is able to infer a time interval that covers for a minimum time period contiguously only small numbers of crossing events, this time interval is assumed to be the red interval.

2. **Single green interval per movement:** With respect to the first assumption, the method assumes that each movement of interest has once in the cycle time a green interval. If the data analysis is not able to infer a single contiguous green interval, the data source is assumed to be too sparse or the movement is in fact controlled in two phases. In such situations, the method is not able to distinguish between both cases, i.e. no green interval estimation can be calculated.

3. **Minimum red and green intervals:** The method assumes furthermore that a movement of interest has a minimum green and therefore a minimal red interval. This minimum time interval is denoted with variables δmin_t and has a duration of at least 7 secs. This is reasoned on a minimum green interval of 5 secs (valid for Germany) [28] and a reasonable amber time of at least 2 secs.

Because frequencies are used to be volatile for small numbers of crossing events, this method proposes to apply a weighted centered moving average on the calculated histogram $h_{cross}(b)$, using parabolic weights. The used principle of weighting tries to reduce influences on a time lag, caused by the smoothing. The smoothed frequencies of the histogram are indicated by the green step line in Figure 4.13 ⓐ. If N specifies the size of the smoothing window and i is the center position in the window, then weight g_j at position j in the window is defined by Equation 4.39. For a window size of $N = 3$, center $i = 1$ and $j = \{0 \dots N-1\}$, weights are $g = [0.75, 1.0, 0.75]$.

$$g_j = 1 - (\frac{j-i}{(N+1)/2})^2 \tag{4.39}$$

In the second processing step ❷, the spectrum of h_{cross} is analyzed in a histogram h_f that counts the number of cycle seconds that belong to a frequency class. The frequency spectrum of h_{cross} is therefore divided into $n_{lvl} = 15$ equidistant classes. This second histogram is depicted in Figure 4.13 ⓑ. Both histograms are used to understand how the crossing events are distributed over the cycle time and over the frequency spectrum.

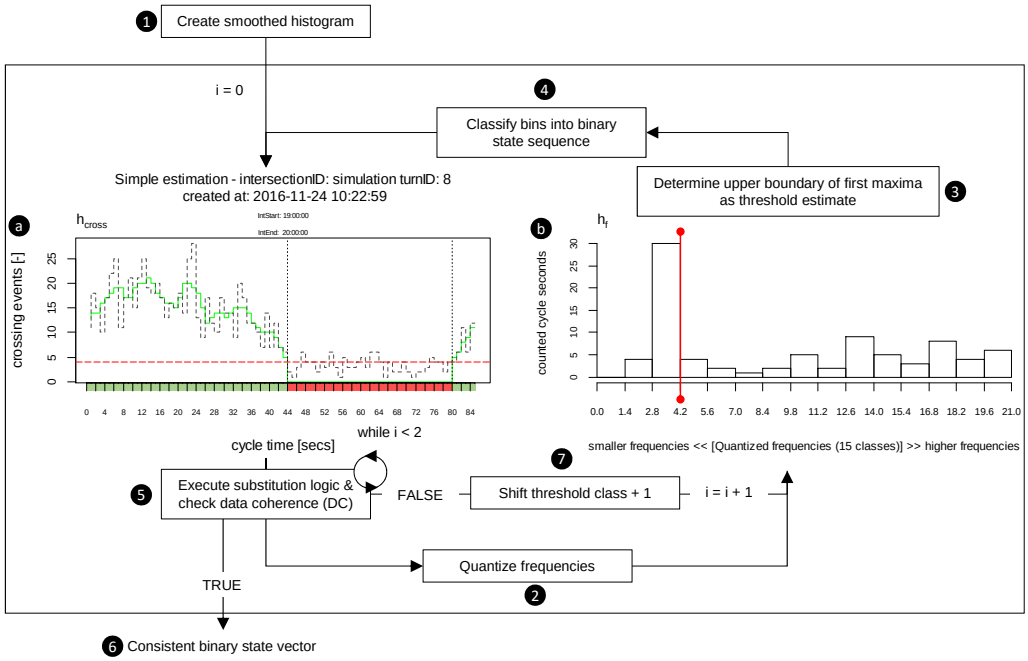


Figure 4.13: Simple green interval estimation process

In the hypothetical case in which no erroneous trajectories are present and crossing events follow a perfectly uniform distribution, h_{cross} would cover a number of contiguous cycle seconds that are free of any crossing events. These seconds represent the red interval. From the perspective of the second histogram h_f, the first bin in $h_f(b_f = 1)$ would have counted these cycle seconds in which no crossing event occurred. Because the number of crossing events is constant over the green interval in a perfectly uniform crossing event distribution, the last bin in $h_f(b_f = n_{lvl})$ would have collected all remaining cycle seconds that do not belong to the red interval. In such an idealistic situation, the second-wise classification of a contiguous red and green interval is trivial. The upper bin boundary of $h_f(b_f = 1)$ can be used as a frequency threshold f_{thr} that needs to be exceeded for each cycle second of h_{cross} in order to obtain a green interval estimate. In a real-world situation with a moderate number of erroneous trajectories, counts in h_{cross} are reaching rarely zero and thus, the first bin in $h_f(b_f = 1)$ contains probably no counts. Instead, a next higher bin in h_f would have collected low frequencies from h_{cross} that occurs relatively often and belongs thus to the red interval.

Hence, the method proposes in the third processing step ❸ to find the first peak in h_f which is taken as a frequency threshold estimate f_{thr}. Exactly this situation is depicted in Figure 4.13 ⓑ. The developed methodology proposes a turning point analysis of a discrete value series to identify peaks and pits within the histogram h_f by applying the algorithm of Grosjean et al. [32]. An exception of this peak selection is the situation, in which the very first turning point in h_f (from the left to right) is a pit. In this special case, the first bin

in $h_f(b_f = 1)$ is used as a frequency threshold. The parameters n_{lvl} and the $\max(h_{cross})$ are scaling the size of δf, which are thus calibration parameters for the methodology.

$$h_f(b_f) = \text{No. of cycle seconds } b \text{ with:} \tag{4.40}$$

$$\delta f \cdot (b_f - 1) \leq h_{cross}(b) < \delta f \cdot b_f, \quad b_f \in \{1 \dots n_{lvl}\} \tag{4.41}$$

$$\delta f = \frac{\max(h_{cross})}{n_{lvl}} \tag{4.42}$$

$$turningpoint \in \{pit, peak\} \tag{4.43}$$

$$turningpoints = [turningpoint_1 \dots turningpoint_n] \tag{4.44}$$

$$f_{thr} = \begin{cases} \delta f \cdot bin(turningpoint_1), & \text{if } turningpoint_1 = peak \\ \delta f, & \text{if } turningpoint_1 = pit \end{cases} \tag{4.45}$$

The selection process of the frequency threshold f_{thr} is described by Equation 4.45, in which the function $bin()$ returns the bin, associated with the first turningpoint. As illustrated in Figure 4.13, the fourth processing step ❹ applies the determined frequency threshold f_{thr} to classify cycle seconds in h_{cross} to a red or a green second within the cycle time. Equation 4.46 assigns each bin of the histogram h_{cross} to its corresponding state, whereas $state$ is used to describe a second-wise state vector that covers the complete cycle time.

$$state = [state_{b=1}, \dots, state_{b=C}] \tag{4.46}$$

$$state_b = \begin{cases} 1, & \text{if } h_{cross}(b) > f_{thr} \\ 0, & \text{otherwise} \end{cases} \tag{4.47}$$

In case of a high number of collected crossing events, counts of consecutive bins $h_{cross}(b)$ and $h_{cross}(b+1)$ are less volatile and the state vector $state$ should contain only one $\overline{peak}$ element, which is a contiguous sequence of ones (i.e. green interval) and likewise a contiguous sequence of zeros (i.e. red interval), denoted as $\overline{gap}$. However, when working with sparsely distributed crossing events, the binary state vector can contain short time periods in which the frequency threshold f_{thr} is not exceeded or vice versa. These situations cause short interruptions and more than one $\overline{gap}$ and $\overline{peak}$ element can occur within the cycle time. The constraint of one contiguous green interval would be violated and no estimation result would be available. As the rejection of an estimation result due to a single second in which not enough crossing events could be counted is not acceptable, a rule-based substitution logic is developed. This logic tries to compensate abundant $\overline{gap}$ and $\overline{peak}$ elements by considering a time constraint $\delta const_t$. The time duration of a $\overline{gap}$ element in the cycle is defined by the function $t(\overline{gap})$. Each $\overline{gap}$ has a start and an end bin that is

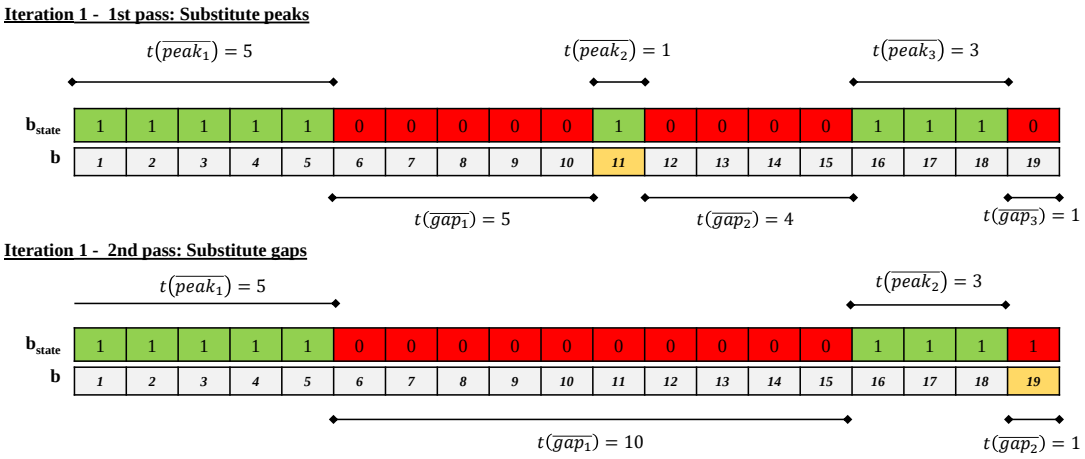


Figure 4.14: Exemplified substitution logic

denoted by $start$ and end. Notation is equivalent for the $peak$ element.

$$\overline{gaps} = \{\overline{gap}_1, \ldots, \overline{gap}_n\} \tag{4.48}$$

$$\overline{gap} = [end, start], \quad state[start, \ldots, end] \stackrel{!}{:=} [0 \ldots 0] \tag{4.49}$$

$$t(\overline{gap}) = (end - start) \mod C \tag{4.50}$$

$$\overline{peaks} = \{\overline{peak}_1, \ldots, \overline{peak}_n\} \tag{4.51}$$

$$\overline{peak} = [end, start], \quad state[start, \ldots, end] \stackrel{!}{:=} [1 \ldots 1] \tag{4.52}$$

$$t(\overline{peak}) = (end - start) \mod C \tag{4.53}$$

The developed substitution logic starts per definition in the 1st pass to search $\overline{peak}$ elements. If more than one peak element is found, its surrounding $\overline{gap_n}$ and $\overline{gap_{n+1}}$ are analyzed. If both gaps are covering a time interval $\delta const_t \geq 2$ secs, the peak will be replaced. In a 2nd pass, the binary state vector is analyzed for gaps. The time intervals of surrounding peak elements are proved under the same constrains and get conditionally substituted. This process repeats till it converges, i.e. the state vector can not be manipulated anymore with respect to the time constraints. See Appendix E for a pseudo code of the used substitution logic.

In the fifth processing step ❺, the data coherence DC of the derived binary state vector is verified. Data coherence covers in total two constraints:

$$DC = \begin{cases} true, & \text{if } (|\overline{peaks}| = 1) \wedge (t(\overline{gap}) \geq \delta min_t) \quad \wedge \\ & (t(\overline{peak}) \geq \delta min_t) \\ false, & \text{otherwise} \end{cases} \tag{4.54}$$

1. Condition $|\overline{peaks}| = 1$ ensures that the binary state vector contains only one peak and is thus a contiguous time interval.

2. Condition $t(\overline{gap}) \geq \delta min_t$ and $t(\overline{peak}) \geq \delta min_t$ ensures a minimum red and green interval with $\delta min_t = 7$ secs.

If the final state vector satisfies these constrains, the output of this sixth processing step ❻ is a contiguous binary state vector and the only remaining peak element $\overline{peak}$ holds the information of start and end of the green interval tg.

$$tg = (start = \overline{peak}_{1,start}, end = \overline{peak}_{1,end}) \tag{4.55}$$

If constraints of data coherence are violated, the procedure is allowed to repeat two times, whereas the threshold f_{thr} ❼ is shifted in each iteration by one bin in h_f. Thus, if after two iterations ($iteration_{nr}$) no binary state vector can be found that satisfy all conditions, the crossing event distribution is assumed to be too sparse or the observed movement gets twice within the complete cycle time a green interval. In such cases the proposed method could fail. As the method up to this stage considers only crossing events, this kind of an estimate is called "simple green interval estimation". The method explained has the following calibration parameters:

Simple green estimation calibration parameters:

- Number of frequency levels: $n_{lvl} = 15$ is the number of bins in h_f used to determine the frequency threshold f_{thr}.
- Required duration: $\delta const_t \leq 2$ secs is the required interval around a $\overline{peaks}$ or a $\overline{gap}$ element to allow its substitution.
- Minimum red and green intervals: $\delta min_t = 7$ secs is the minimum duration of a green or red interval.
- Number of allowed iterations: $iteration_{nr} = 2$ is the number allowed iterations to find a coherent time interval. In each iteration f_{thr} is increased by one bin in h_f.

Simple Green Interval Estimation at Fixed-Time Intersections

A simulation study has been established in this thesis to make a prove of concept and to explore effects of different sample sizes and degrees of saturation at a through movement of a signalized intersection. The degree of saturation x is thereby defined as the quotient between the traffic flow q and the available capacity C at the movement. The very same objectives have been also analyzed by Axer et al. [6] with the help of an earlier development state of the green interval estimation methodology. The author assumes that at movements with a very low saturation, nearly no tailback occurs and thus a large proportion of the traffic demand utilize intensively only the first seconds of the provided green interval. This assumed phenomenon could cause an

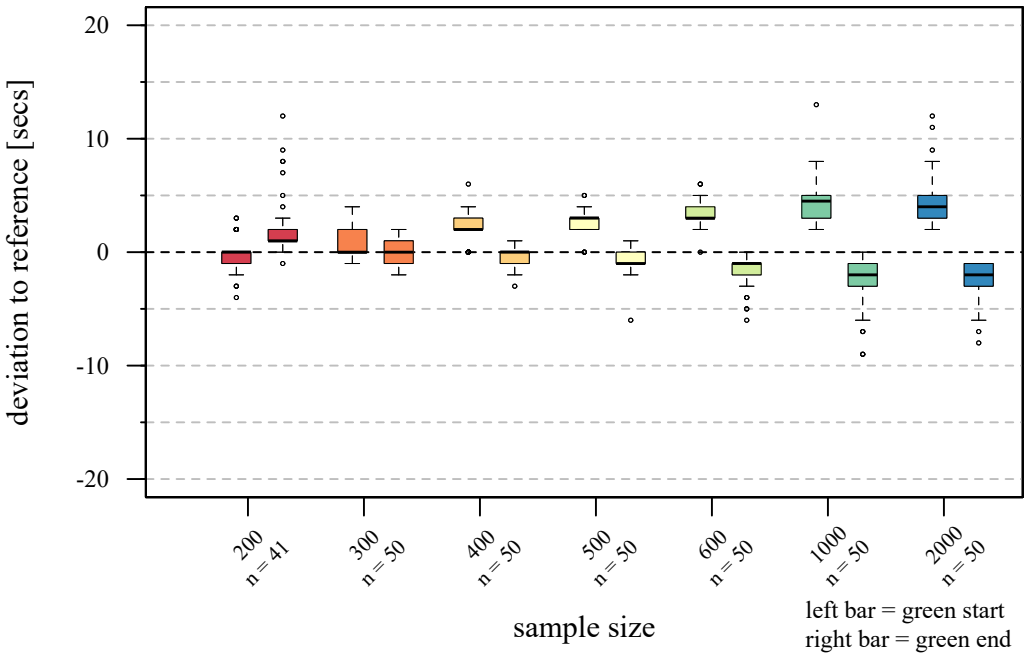
underestimated green end because the methodology requires trajectories that cover the complete green interval.

Furthermore, this study should investigate the typical resulting variability of the introduced simple green interval estimation when providing the algorithm different trajectory volumes. These volumes contain thereby randomly distributed error effects that can occur likewise in real-world scenarios. It should be remarked at this point that a simulation study is never able to prove all imaginable combinations of intersection layouts, demand characteristics and signal control strategies. To prove the general real-world applicability of the developed methods, this thesis refers to the later results of the real-world study, summarized in Chapter D. The previously introduced calibration parameters have been used and the simulation study has been set up as follows.

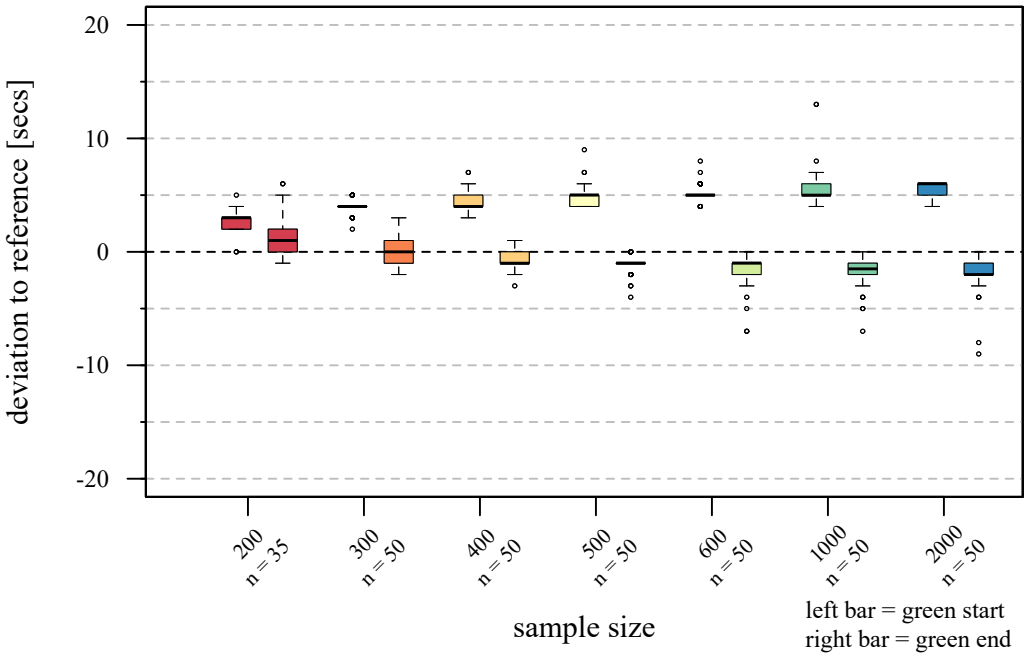
- The signal control strategy in this simulation is fixed-time, with a cycle length of $C = 85$ secs and a green interval that covers the cycle seconds 1 to 63. The amber interval is 2 secs and is counted as green. The long green interval is used to get pessimistic estimates of the needed sample size to infer a contiguous green interval.
- Two demand scenarios are simulated. Each scenario contains a total population of 50,000 trajectories that arise from 30 replications per scenario. In the first demand scenario (1st: Scenario) movement's degree of saturation is $x \approx 0.5$, whereas in the second demand scenario (2nd: Scenario) movement's degree of saturation is $x \approx 0.05$. These scenarios are used to explore effects on different crossing event distributions.
- Two types of errors are emulated in both scenarios. First, a GPS error $E_{GPS,k}$ for each trajectory k has been applied (see Equation 4.15 in Section 4.3). Second, a proportion $e_{rate} = 0.2$ of the input data are random trajectories, used to prove the robustness of the developed frequency threshold f_{thr}.
- The simulation supplies the estimation process different sample sizes that range from 100 to 2,000 trajectories. Each $sample_{lvl}$ already contains the random error rate $e_{rate} = 0.2$.
- As the emulated error influences are stochastic effects, the estimation procedure is replicated 50 times for each $sample_{lvl}$. These 50 replications are used to make findings regarding the result variability of the simple green interval estimation for a given $sample_{lvl}$.

The experimental design and the simulated trajectory volumes should appear and act like a real-world FCD source that is capable to generate the mentioned trajectory volumes for a yearly workday aggregate and a time interval (e.g. 1 hour) for which signal timing parameters are tried to be estimated. The cycle length is from the perspective of the estimation methodology unknown, which means that prior to the green interval estimation, the cycle

length needs to be inferred. The following results focus only on the green interval estimation because Section 4.4 has already presented findings of data requirements to estimate the cycle length information.



(a) Error distribution of green interval for 1st: Scenario($x \approx 0.5$)



(b) Error distribution of green interval 2nd: Scenario($x \approx 0.05$)

Figure 4.15: Simulation-based deviation analysis - simple estimation

Figure 4.15a illustrates for the first scenario the deviation between the reference and the estimated green intervals separately for the start and end of green in a box-plot. A negative deviation means, the estimated green start or end is delayed to its corresponding reference value. For low sample sizes of approximately $sample_{lvl} = 100$, constraints of data coherence are never satisfied and thus no estimates are available. Towards first expectations the results for the start and end of green are slightly better for smaller samples sizes. Furthermore, it is obvious that an increasing sample size cause likewise an systematic offset at the start and end of green. Both phenomena can be explained by the influence of the trajectory interpolation and the applied random error rate e_{rate}.

Because the filtering of erroneous crossing events (see Section 4.3) is not perfect, higher samples sizes cause an error cumulation especially close to the green start till a level is reached on which the estimation results are affected. This finding is supported by Figures C.1.1 and C.1.2 in Appendix C.1 that depict the crossing event distributions for a moderate ($sample_{lvl} = 500$) and very high number ($sample_{lvl} = 2000$) of analyzed trajectories. Even the linear influence of the random error rate starts first to affect the estimation results substantially only when reaching higher samples sizes. An example calculation demonstrates the interdependencies. At a sample size of 200 trajectories only 40 erroneous crossing events are spread uniformly over the cycle time. In contrast thereto, when analyzing 500 trajectories already 100 erroneous random crossing events interfere the simple green estimation process. As the distribution of erroneous events is a stochastic process, it can happen that some erroneous events are taking place at the very same cycle second, causing an error agglomeration. To be able to find despite this agglomeration a contiguous green interval, the estimation methodology needs to increase the frequency threshold f_{thr}. When reaching sample sizes of approximately $sample_{lvl} = 1000$, the deviation between estimate and reference converges in this simulation study.

The 25th and 75th percentiles in Figure 4.15a allow to conclude the tendency that in the first scenario the green start is estimated approximately 3-5 secs too early, whereas the end of green is approximately 1-3 secs delayed. Thus, this very long green interval requires approximately 200-300 trajectories to allow the inference in nearly all 50 replications. Even for 200 supplied trajectories the constraints of data coherence are not always satisfied and thus not all of the 50 replications led to a green interval estimate.

Nearly the very same findings hold even for the second demand scenario ($x \approx 0.05$) in which the interquartile range of errors are slightly smaller, whereas the absolute error for the same sample size seems to be slightly larger. This effect needs to be related to the different degrees of saturation and thereof caused tailback scenarios. The occurrence of tailbacks needs to influence the number of vehicles that start their acceleration very close to the stop bar. Section 4.2 has shown that vehicles, which are crossing the stop

bar slowly during their acceleration phase could cause erroneous crossing events. Thus, it is reasonable to assume that at higher degrees of saturation, most vehicles are starting their acceleration phase some meters apart from the stop bar and the crossing speed is relative high when reaching the stop bar. The number of erroneous crossing events caused by the linear trajectory interpolation should be therefore lower for higher degrees of saturation and vice versa. This argumentation is supported by Figures C.1.2 and C.1.4 in Appendix C.1. Both figures show the output of the green interval estimation for a trajectory sample size of $sample_{lvl} = 2000$ for the two tested scenarios. It can be proved that at the lower saturated movement crossing events are strongly agglomerated in the first 8 secs of the cycle time. Moreover, the lower saturated movement indicates indeed much more erroneous crossing events in advance of the true green start, which takes place in the first cycle second.

Nevertheless, if a FCD source is able to provide in total the same number of trajectories for both degrees of saturation, the expected estimation error range for this tested fixed-time controlled intersections are relative similar. It can be stated as a conclusion that the simple estimation methodology suffers mainly from a too early detected start of green. This method and data inherent error has a range of 3-5 seconds. The estimated end of green shows in both scenarios a smaller error, whereas the estimates are mostly delayed with 1-3 secs. The mentioned values represent the 25th and 75th percentiles.

4.5.2 Combined Green Interval Estimation

Till this step, the proposed green interval estimation methodology exploits only crossing events at a single movement. The established simulation study has demonstrated especially the negative effect of the linear trajectory interpolation that cause mainly a too early estimated green start. As already introduced in Section 4.2, the green interval estimation process can also exploit information from stopped vehicles. Collected and counted stop seconds of trajectories can be interpreted as a complement of crossing events. Under perfect circumstances, e.g. a low sampling interval $t_d = 1$ sec and a fixed-time controlled intersection, histograms of both datasets should not overlap each other. However, when trajectories are sampled with a sampling interval of $t_d = 15$ secs, both histograms need to overlap. The author assumes that the complementary usage of crossing events and stop seconds can be beneficial and therefore lead to the following advantages:

1. **Compensation of the interpolation error:** If the number of crossing events and measured stop seconds are well leveled, too early estimated green starts and delayed estimates of the green end could be compensated.

2. **Freedom of contradiction:** A simple green interval estimate is not verified against collected stop seconds of the very same movement. Thus, if the histogram of stop seconds is subtracted from the histogram of crossing

events, a particular number of crossing events per cycle second remains that is free from any contradiction. The remaining and therefore verified number of consistent crossing events per cycle second needs likewise to satisfy the constraints of data coherence to infer a plausible green interval.

$$s_k(b) = \begin{cases} 1, & \text{if} \quad T_{state,C_{best},k}(b) = stopped \\ 0, & \text{else} \end{cases} \tag{4.56}$$

$$h_{stop}(b) = \sum_{\forall k} s_k(b) \cdot w_k(b) \tag{4.57}$$

The combined green interval estimation methodology relies again on a histogram analysis. The definition of the stop second histogram h_{stop} is very similar to the previous histogram, used to described frequencies of crossing events over the cycle time. In contrast thereto, each stop second s_k at bin b is weighted with respect to its position in the tailback w_k (see Section 4.2.2). Weighting takes into consideration that vehicles close to the stop bar are assumed to generate a more reliable estimate of the red interval than vice versa. All methodical steps of the combined green interval methodology are very similar to the simple green estimation procedure. Necessary processing steps are explained step-by-step in the following paragraphs, whereas each step is depicted in Figure 4.16.

In a first processing step ❶, both histograms h_{stop} and h_{cross} are analyzed by the already established simple estimation procedure. This classification process leads to two binary state vectors $state_{cross}$ and $state_{stop}$. Indicies $cross$ and $stop$ are used to distinguish their relationship to the histograms. These state vectors are analyzed and the substitution logic tries to compensate abundant $\overline{gaps}$ and $\overline{peaks}$ with respect to the time constraint $\delta const_t$. If one out of two state vectors violates constrains of data coherence, no combined estimation result can be calculated and the method fails.

In a second processing step ❷, histograms h_{stop} and h_{cross} are matched with their coherent binary state vectors to tolerate only bins b that satisfy the constraints of data coherence. Equation 4.58 describes this procedure formally which is also illustrated in Figure 4.16.

$$\hat{h}_{cross}(b) = \begin{cases} h_{cross}(b), & \text{if } state_{cross,b} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.58}$$

$$\hat{h}_{stop}(b) = \begin{cases} h_{stop}(b), & \text{if } state_{stop,b} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.59}$$

In a third processing step ❸, the stop second histogram $\hat{h}_{stop}$ is weighted with the factor w_{red}. This factor is needed because stopped vehicles are collected from a space interval that covers an area of $\delta s_{stop} = 35$ m upstream the stop bar. Collecting data from a space interval has the potential to gather in

the same time period more stop seconds than crossing events per cycle second. Thus, in case of regular and long tailbacks counted stop seconds can dominate massively the number of crossing events per cycle second and the error influence of trajectory interpolation close to the green start is overcompensated. Overcompensation means in this context that the detected start of green is shifted from the true second to a later second in the cycle time. The weighting factor w_{red} is thus used as a method to keep both data sources leveled. Because its fundamental influence, weighting parameter w_{red} is considered to be one of the most important calibration parameters.

In the fourth processing step ❹, the stop second histogram $\hat{h}_{stop}$ is subtracted from the crossing events histogram $\hat{h}_{cross}$ (see Figure 4.16) and the remaining counts per cycle second are stored in a new histogram $\hat{h}_{rem}$. The remaining number of consistent crossing events over the cycle time is thereby indicated with the help of black dashed stepline.

The last fifth processing step ❺ has the objective to derive a new second-wise state vector $state_{rem}$ from $\hat{h}_{rem}$ according to Equation 4.60 that needs to satisfy again constrains of data coherence. The substitution logic is again used to compensate abundant $\overline{gaps}$ and $\overline{peaks}$. In case of a consistent state vector $state_{rem}$, the only remaining peak element ($\overline{peak_1}$) contains analogously to the previous simple estimation procedure (see Section 4.5.1) the start and end of the green interval.

The overall concept of the combined green interval estimation is straightforward, because the method determines the green interval based on the remaining crossing events per cycle second and constrains of data coherence ensures the consistency of the second-wise state vector $state_{rem}$. If the substitution logic is not able to infer a time interval that satisfy all constraints of data coherence, this method fails and the signal parameter estimation falls back to the result of the simple estimation procedure. As this method takes crossing events and stop seconds into account, the method is called "combined green interval estimation".

$$h_{rem} = \hat{h}_{cross} - w_{red} \cdot \hat{h}_{stop} \tag{4.60}$$

$$state_{rem} = \begin{cases} 1, & \text{if } h_{rem,b} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.61}$$

Combined green estimation calibration parameters:

- Number of frequency levels: $n_{lvl} = 15$ is the number of bins in h_f used to determine the frequency threshold f_{thr}.
- Required duration: $\delta const_t \leq 2$ secs is the required interval around a $\overline{peaks}$ or a $\overline{gap}$ element to allow its substitution.
- Minimum red and green intervals: $\delta min_t = 7$ secs is the minimum duration of a green or red interval.

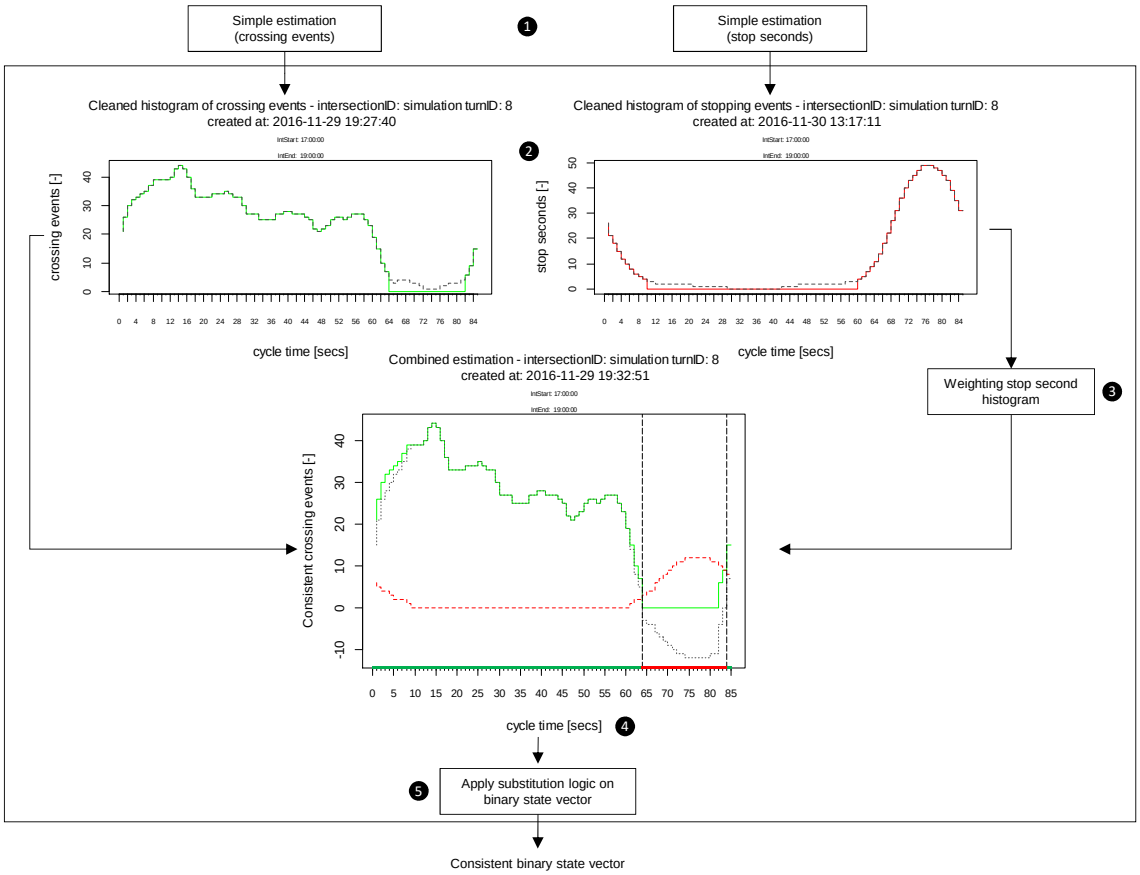


Figure 4.16: Combined green interval estimation process

- Number of allowed iterations: $iteration_{nr} = 2$ is the number of allowed iterations to find a coherent time interval. In each iteration f_{thr} is increased for one bin in h_f.
- Weighting factor: $w_{red} = 1/4$ is a factor that tries to level histogram of stop seconds and crossing events with the objective to compensate effects of the interpolation error.

Combined Green Interval Estimation at Fixed-Time Intersections

The combined green interval estimation methodology has been analyzed according to the same simulation scenarios that have been used to evaluate the simple green interval estimation procedure (see Subsection 4.5.1). For the first scenario ($x \approx 0.5$), the accuracy of the green interval estimates is increased (see Figure 4.17) and the systematic offset at the green start is well compensated by the complementary use of the stop seconds. For small sample sizes, lower than 200-300 trajectories, constraints of data coherence often lead to an estimation result rejection, whereas most of the remaining results, indicate a green start and a green end error less than 5 secs.

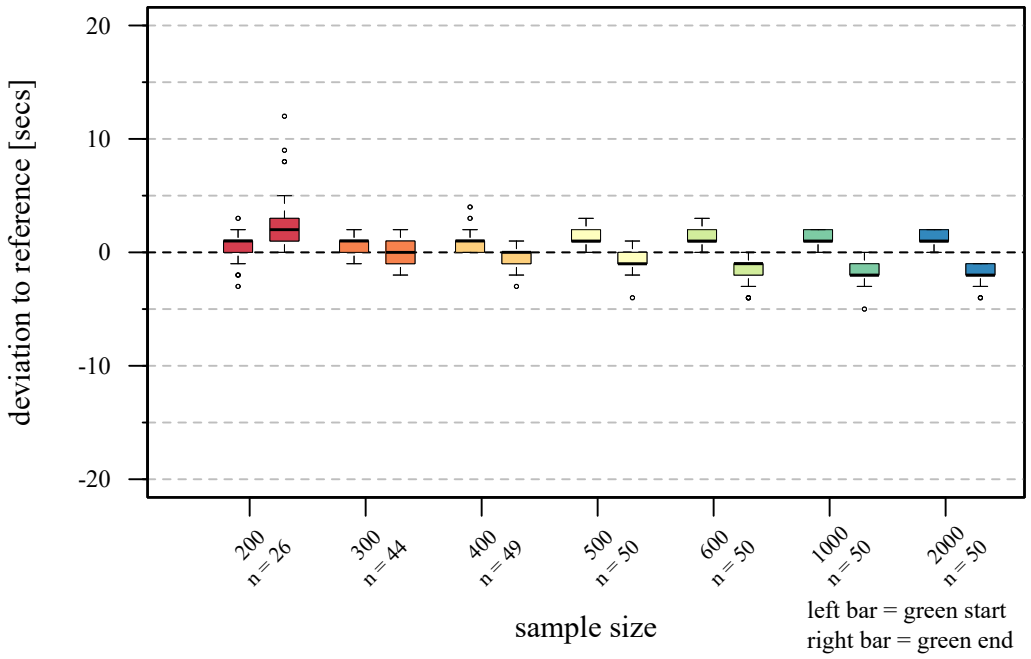


Figure 4.17: Error distribution of green interval for 1st: Scenario($x \approx 0.5$)

More interesting are the results of the very low saturated movement that has been simulated (see Figure 4.18). A comparison between the results of the combined and simple estimation procedure (see Figure 4.15b) reveals nearly no substantially differences. Thus, errors of the green start and end are in the very same error range and show the same offsets. A more detailed data analysis allows to give some suggestions of potential reasons.

First, on low saturated movements less tailbacks occur and thus, less stop seconds can be detected by the trajectories. If the total number of stop seconds is too low, the positive effect of a compensated interpolation error does not emerge. The problem is intensified by the parameter w_{red}. This parameters is a static and global weight which has the objective to level the number of measured stop seconds. In all simulation studies this factor has been set to $w_{red} = 1/4$, which seems only to be legit for medium saturated intersections, with $x \approx 0.5$. Thus, in the special case of a very low saturated movement, it is necessary to increase w_{red} to achieve again a compensation of the interpolation error. The very best solution would be, to estimate reliably movements' degree of saturation and to adapt w_{red} automatically. Such a reliable estimation procedure could not be developed and systematically calibrated in this thesis. Thus, $w_{red} = 1/4$ is in the actual implementation of the signal parameter estimation a static factor, mainly valid for medium degrees of saturation.

Second, because tailbacks are in tendency much smaller on low saturated movements, vehicles are starting to accelerate mainly close to the stop bar with low speeds. On the opposite, if tailbacks are longer, a larger number of vehicles will start to accelerate far away from the stop bar and thus vehicles

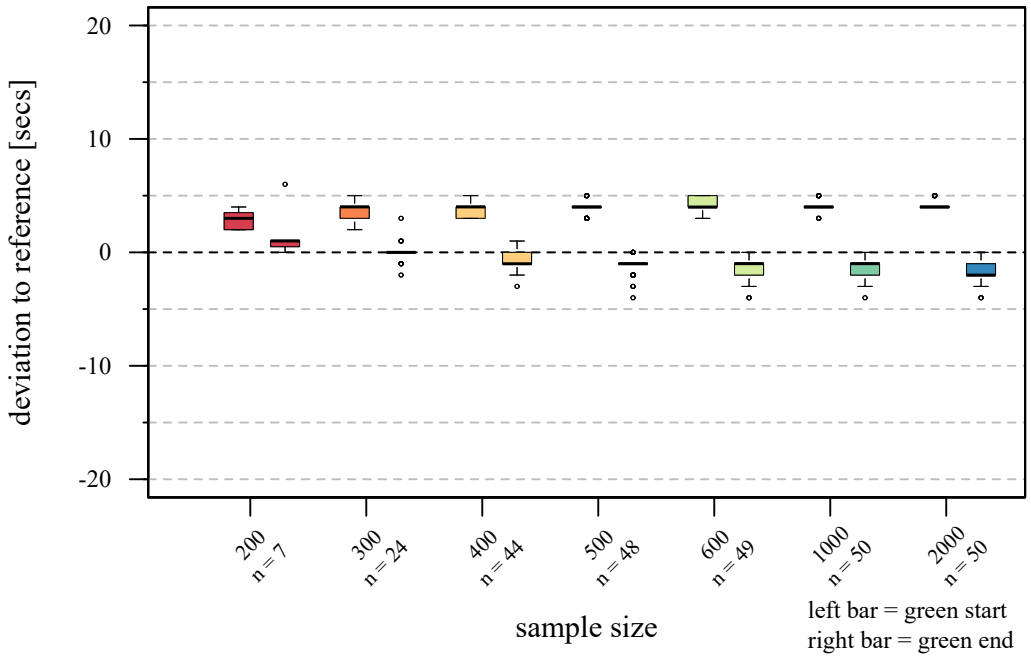


Figure 4.18: Error distribution of green interval for 2nd: Scenario($x \approx 0.05$)

cross the stop bar with higher speeds and a smaller interpolation error. This interdependency has already been introduced in Subsection 4.5.1. All in all, the second scenario is affected by two negative aspects which are superimposing the simulated trajectories. To support this hypothesis, Figures C.1.5 and C.1.6 in appendix C.1 illustrate result plots of the combined estimation procedure that demonstrate the distributions of the crossing events and stop seconds over the cycle time.

It could be concluded that the combined estimation methodology has the potential to compensate almost completely the negative effect of the trajectory interpolation. A successful compensation requires therefore the correct and automated calibration of the weighting factor w_{red} in order to achieve good estimation results for different degrees of saturation at individual movements.

Green Interval Estimation at Intersections with Vehicle Actuated Control

Up to this subsection, the thesis presented important findings for the simple and the combined green interval estimation tested on different simulated fixed-time controlled intersections. These findings are very valuable and allow to assess the method inherent result variability, if the very same intersection is analyzed by different randomly selected data subsets. This section refers to results of the research from Axer et al. [5] that had the objective to explore the green interval estimation by considering a simple traffic actuated signal control with a fixed cycle length. Because traffic actuated intersections are known to react towards high traffic demand like fixed-time controlled [77, 60],

this study has also the objective to figure out, how the green interval estimation methodology is reacting to this phenomenon. Furthermore, this study wants to assess the potential of a traffic signal state prediction when using a static set of estimated signal timing parameters (t_{ref}, cycle length C, green interval tg) to predict future signal states for the simulated traffic light in a time period, in which the degree of saturation of a movements is $x \approx 1.0$. To be able to quantify exclusively the influences of a traffic actuation, a microsimulation has been set up as follows.

- The signal control has two phases and a constant cycle length of $C = 85$ secs. The green time for the movement of interest starts at second 85 and is fixed, earliest green end is at second 30, latest green end is at second 42. The maximum available green interval is therefore 43 seconds. Signal state amber is handled again as green. The signal phase is interrupted in the simulation, if the vehicle time headways are greater or equal to 5 secs and the alternative phase is requested.
- The mean degree of saturation on the movement of interest is $x \approx 1.0$.
- Two types of errors are emulated in both scenarios. First, a GPS error $E_{GPS,k}$ for each trajectory k has been applied (see Equation 4.15 in Section 4.3). Second, a proportion of the input data are random trajectories $e_{rate} = 0.2$, used to prove the robustness of developed frequency threshold f_{thr}.
- The simulation supplies the estimation process different sample sizes ranging from 100 to 600 trajectories. Each $sample_{lvl}$ already contains the random error rate $e_{rate} = 0.2$.
- As the generated errors are stochastic effects, the estimation procedure has been replicated 50 times for each $sample_{lvl}$ to make again findings of the method related result variability.

Table 4.8 presents the results of this study. Up to a sample size of about 200 trajectories 0 out of 50 replications (rep) are able to generate an estimate. Thus, the insufficient trajectory volumes and therefore deduced estimates do not satisfy the constraints of data coherence. For higher sample sizes, the result variability of the estimated green start are very similar to the previous simulation studies. The mean (mn) green start in Table 4.8 shows a systematic offset that is increased for higher sample sizes. As already concluded, this effect can essentially be compensated by the application of the combined estimation methodology. In this simulation study, the usage of the combined estimation methodology causes an overcompensation of the interpolation error and thus, the mean green start is shifted approximately 3 secs away from the true green start, which is at cycle second 85 (see Table 4.8). The phenomenon is plausible because the weighting factor w_{red} is not changed in this traffic

actuated scenario and does not seem to fit well with movement's degree of saturation, which is about $x \approx 1$. At the nearly oversaturated movement, the number of collected stop seconds dominates the crossing events. This hypothesis is supported by Figures C.3.9 and C.3.10 in this Appendix C.3.

Especially the green end estimates need to be underlined in this simulation study. Simple and combined estimation methodology have always detected for different samples sizes the maximum available green interval, specified for this traffic actuated movement. Thus, it can be remarked that the high degree of saturation has forced the signal control to use the maximum available green interval often. To prove this hypothesis, the study has used the estimated signal timing parameters, green start $tg_{start} = 2$, green end $tg_{end} = 42$ and the cycle length of $C = 85$ secs to predict signal states of one replication in the simulation.

$$tg = [start, end], \quad state[b = start, \ldots, b = end] \stackrel{!}{:=} [1 \ldots 1] \tag{4.62}$$

$$bin(t_{act}) = (t_{act} - t_{ref}) \mod C \tag{4.63}$$

This experiment emulates the case in which the developed estimation methodology would have been used to infer the signal timing parameters from a historical trajectory dataset during the peak hour and the same static parameter set is applied to predict future signal states within the very same hour. The prediction is straightforward, if the reader remembers the definition of an estimated green interval which has been described in Section 4.5.1 as second-wise state vector $state$. Each element $state_b$ indicates its signal state over the cycle time. Thus, it needs to be determined, which bin of the cycle time

Table 4.8: Result variability - Simple estimation methodology

$sample_{lvl}$	rep	$mn(tg_{start})$	$mn(tg_{end})$	$sd(tg_{start})$	$sd(tg_{end})$
100	0	-	-	-	-
200	50	0.60	41.52	1.36	1.06
300	50	84.54	42.74	1.47	1.25
400	50	83.42	44.27	1.25	2.03
500	50	82.94	44.89	1.21	1.93
600	50	82.40	45.74	1.10	2.51

Table 4.9: Result variability - Combined estimation methodology

$sample_{lvl}$	rep	$mn(tg_{start})$	$mn(tg_{end})$	$sd(tg_{start})$	$sd(tg_{end})$
100	0	-	-	-	-
200	50	1.98	41.24	0.93	1.07
300	50	2.06	41.58	0.76	1.11
400	50	1.88	41.88	0.74	0.84
500	50	1.88	42.04	0.59	0.72
600	50	1.74	42.16	0.59	0.73

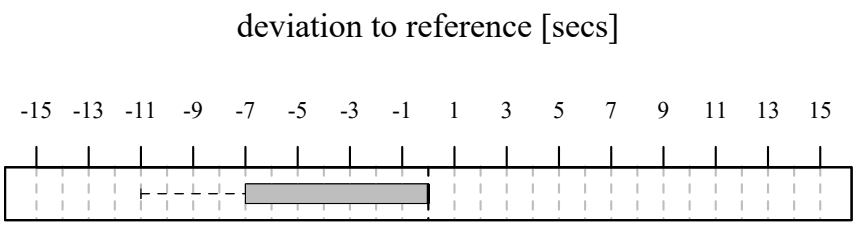


Figure 4.19: End of green prediction error, $tg_{end} = 42$ sec, $n = 170$

is associated with a actual or future time point t_{act} by applying Equation 4.63. A short calculation demonstrates this procedure for a set of fictional signal timing parameters. The time calculation in Equation 4.64 uses the UNIX (Uniplexed Information and Computing System) time in which the timestamp "1970-01-01 00:00:00 (UTC)" is equal to zero [75].

$$C = 90 \text{ secs}, \; t_{ref} = \text{1980-01-01 00:00:00 UTC} \tag{4.64}$$

$$tg = [start = 15, end = 37], \quad state[state_{b=15}, \ldots, state_{b=37}] \overset{!}{:=} [1 \ldots 1]$$

$$bin(t_{act}) = (\text{2016-12-16 10:16:10 UTC} - \text{1980-01-01 00:00:00 UTC}) \mod 90$$

$$bin(t_{act}) = 1166350570 \text{ secs} \mod 90 \text{ secs} = 70 \text{ secs}$$

$$state_{b=bin(t_{act})} = 0$$

The box-plot in Figure 4.19 illustrates the deviation in seconds between the estimated and ground truth green end, whereas the ground truth is derived by logging signal states second-by-second in the simulation. Figure 4.19 and box-plot median position at second zero allows to conclude that the high traffic demand forces the actuated signal control to use the maximum specified green interval in 50 % of 170 observed cycles in the simulation. This means that in 50 % the estimated green end matches perfectly the true value. The traffic demand in this simulated example were not stationary and thus, some cycles were adapted to the stochastic traffic demand changes. This experiment underlines the importance of similar traffic demands for the same time of day in order to be able to predict the traffic signal states reliably even for simple traffic actuated intersections. If the real-world mean traffic demand is not nearly stationary for a particular hour of the weekday, estimated traffic signal states would be only a very rough estimator of the true green interval.

4.6 Result Clustering and Time Assignment

4.6.1 Concept

The described estimation procedure allows to deduce signal timing estimates for a given hour h of the yearly workday or weekend period for an individual movement m of interest. The simulation studies have shown that the critical trajectory volume to satisfy constrains of data coherence and thus to allow

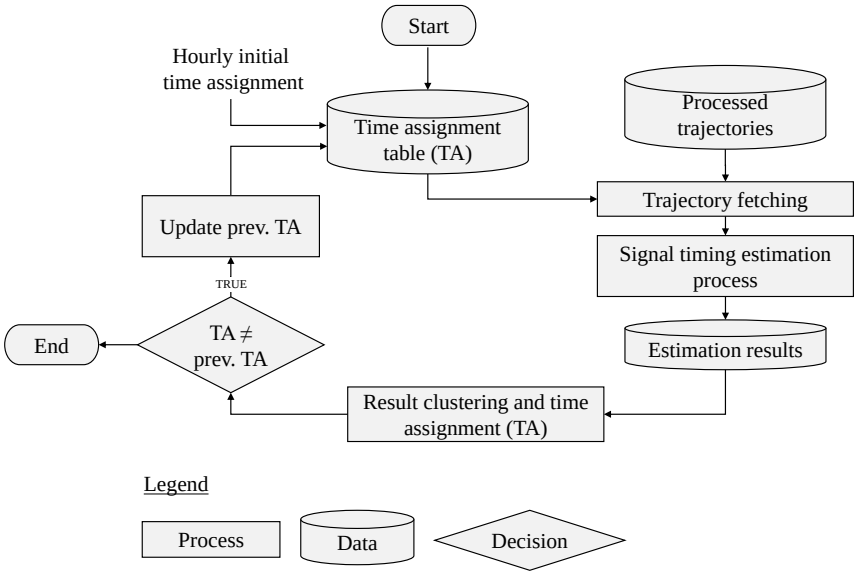


Figure 4.20: Result clustering and time assignment approach

green interval estimation within a particular hour of the workday period is approximately about 200-300 trajectories. As these volumes have been derived by a simulation analysis of single signal programs, findings should be interpreted as a rough indicator of the needed samples sizes to allow a robust green interval estimation. There may be situations with longer cycle lengths and green intervals that require more trajectories to overcome the constraints of data coherence which have not been simulated in this thesis.

As motivated in Subsection 1.2, this thesis has also the objective to develop a verification procedure that takes green estimates of conflicting movements into account. The desired functionality requires therefore estimates on all conflicting movements for a given hour. Thus, it is important to think on the problem that a result verification could fail because the trajectory volumes on other movements are insufficient and do not allow the estimation of green intervals. Furthermore, it is well known that commuter traffic leads typically to a very asynchronous traffic demand in our road networks. While during the morning rush hour high volumes of trajectories are available inbound to metropolitan areas, on the outbound direction collected trajectory volumes might be insufficient. This phenomenon is naturally reversed on the evening rush hour.

Motivated by this problem, the thesis presents in this section an approach that tries to learn hours in which an intersection reveals a similar signal timing parametrization, at least on one analyzed movement. If these similarities are observable, the method proposes to cluster these hours. The cluster and its associated hours are stored in a time assignment table (TA) which is used to provide the signal timing estimation process an more aggregated dataset of all clustered hours. The aggregated hours may contain enough

trajectories in order to satisfy constrains of data coherence even on minor movements with smaller trajectory volumes. The flowchart in Figure 4.20 illustrates this iterative procedure. The described method makes the following two assumptions:

1. **Clustering of similar green intervals:** If a movement shows at least in two hours green interval estimates within a certain tolerance and the cycle lengths are equal, the method concludes that these hours are controlled by a signal program in which parameters of signalization for the complete intersection are in mean relatively constant.
2. **Minimizing the amount of clusters:** As each movement m has at this stage of the signal timing estimation process individual estimates for single hours, clusters of similar timing parameters would be found on each movement m separately. Thus, a cluster selection process is needed that tries to minimize the total amount of clusters at the whole intersection.

Before introducing the details of the clustering process, a definition of the required clustering tolerance $\delta t_{cluster}$ has to be found. The thesis considers therefore results from previously introduced simulation studies (see Subsections 4.5.1 and 4.5.2).

These studies have shown that even on fixed-time controlled intersections, the estimated start and end of green seems to have a certain variance which could visually inferred from boxplots by examining the interquartile range between 25th and 75th percentile. This range seems to be relatively constant which supports the assumption that the result variance is inherent for the used method of green interval estimation. In tendency, very high sample sizes and therefore a better data provisioning of more than 400 trajectories reduces this variance, at least in simulations. In case of the simple estimation procedure, the interquartile range reaches in maximum 3 secs while the combined estimation shows a maximal interquartile range of 2 secs. However, it is more than reasonable to assume that results, derived from simulations, are always superior to the real-world and thus $\delta t_{cluster}$ is set to 4 secs in this thesis. The variable $\delta t_{cluster}$ is another important calibration parameter of the developed signal timing estimation methodology. The example in the following explains the developed cluster procedure step-by-step for a single cycle length as the occurrence of multiple different cycle lengths within the workday period would only increase the complexity in this example.

The clustering of similar green starts and ends is based on two distance matrices $M_{tg_{start}}$ and $M_{tg_{end}}$ that describe separately the absolute time distance in seconds between all green starts tg_{start} and ends tg_{end} for all hours h on a single movement m for which estimates are available. $M_{tg_{start}}$ is therefore the distance square matrix that holds for each index pair i, j the absolute time distance between all green starts $m_{tg_{start,i,j}}$, measured in seconds. The definition of $M_{tg_{end}}$ is analog to $M_{tg_{start}}$, whereas the start of green is substituted by the end of green.

$$tg_{start} = [tg_{start_1}, tg_{start_i} \ldots, tg_{start_n}] \tag{4.65}$$
$$M_{tg_{start}} \in \mathbb{N}^{n \times n}, M_{tg_{end}} \in \mathbb{N}^{n \times n}, n = |h|, h \subset H = \{0 \ldots 23\} \tag{4.66}$$
$$m_{tg_{start,i,j}} = min\{|tg_{start,i} - tg_{start,j}|, \tag{4.67}$$
$$|tg_{start,i} + C - tg_{start,j}|, \tag{4.68}$$
$$|tg_{start,j} + C - tg_{start,i}|\} \tag{4.69}$$

Clustering is allowed as long as $m_{tg_{start,i,j}}$ and $m_{tg_{end,i,j}}$ are simultaneously less than or equal to $\delta t_{cluster} = 4$ secs. The implemented clustering process utilizes the complete-linkage method [23] to ensure that the maximum distance between all objects of the same cluster are within the required time tolerance. The result clustering is applied for each movement and results are stored in the set cls. Each cluster $cl_{m,nr} \in cls$ contains multiple attributes. Index m denotes the movement membership and index nr denotes the number of the found clusters in the movement. Information $L_{m,nr}$ is the count of clustered hours, whereas $l_{m,nr}$ describes the count of interconnected hours for the same cluster. Interconnected means in this context that all hours are adjacent.

Once all clusters are stored in cls, a selection process determines the best available cluster in cls, denoted with cl_{best}. The selection for this best cluster starts by searching the cluster with the largest count of interconnected hours. If two clusters have the same counts of interconnected hours $l_{m,nr}$, the selection process maximizes the number of similar but not connected hours $L_{m,nr}$.

This concept is demonstrated in Figure 4.21 which shows row-wise estimation results of two movements $m = \{2, 8\}$ for multiple hours. The different colors are used to distinguish the found cluster structure at each movement. The cluster $cl_{8,2}$ indicates the highest count of connected hours. Thus, this cluster $cl_{8,2}$ is selected and inserted into the set of best clusters cls_{best}. Next, the estimation results on both movements that belong to the hours of best found clusters $cl_{8,2}$ are deleted for each movement.

Movement (m = 2)

h	C	tg_{start}	tg_{end}	$cl_{m,nr}$	
4	90	33	63	$cl_{2,1}$	❷
5	90	32	61	$cl_{2,1}$	❷
6	90	35	60	$cl_{2,1}$	❷
7	90	31	58	$cl_{2,2}$	❸
8	90	60	3	$cl_{2,3}$	
9	90	62	1	$cl_{2,3}$	
10	90	63	87	$cl_{2,4}$	
11	90	59	2	$cl_{2,5}$	
12	90	60	77	$cl_{2,6}$	❹

Movement (m = 8)

h	C	tg_{start}	tg_{end}	$cl_{m,nr}$	
4	90	32	61	$cl_{8,1}$	
5	90	30	59	$cl_{8,1}$	
6	90	NA	NA	NA	
7	90	32	59	$cl_{8,1}$	
8	90	61	2	$cl_{8,2}$	❶
9	90	63	4	$cl_{8,2}$	❶
10	90	59	1	$cl_{8,2}$	❶
11	90	60	2	$cl_{8,2}$	❶
12	90	60	75	$cl_{8,3}$	

Time assignment

h	$cl_{m,nr}$	
4	$cl_{2,1}$	❷
5	$cl_{2,1}$	❷
6	$cl_{2,1}$	❷
7	$cl_{2,2}$	❸
8	$cl_{8,2}$	❶
9	$cl_{8,2}$	❶
10	$cl_{8,2}$	❶
11	$cl_{8,2}$	❶
12	$cl_{2,6}$	❹

❶ Iteration counter for best cluster selection

Figure 4.21: Principle of result clustering

The described procedure is repeated for all remaining estimation results until all hours in which estimates have been originally available are deleted. Figure 4.21 shows with the circles ❶ to ❹ the best selected cluster in each iteration. At the end, the procedure has selected a set of clusters that tries to minimize the total number of necessary clusters to cover all hours. Each best cluster in cls_{best} stores its corresponding hours used to derive a new time assignment table (TA). Equation 4.70 presents the result for the introduced example. A pseudo code of this cluster selection procedure is described in Appendix E.

$$cls_{best} = \{cl_{8,2}, cl_{2,1}, cl_{2,2}, cl_{2,5}\} \tag{4.70}$$

$$TA = \{\{8, 9, 10, 11\}, \{4, 5, 6\}, 7, 12\} \tag{4.71}$$

4.6.2 Concept Verification

Different simulation studies have already proved the correct functionality of the cycle length and the green interval estimation methodology. Thus, the next paragraphs have the objective to prove the correct functionality of the developed clustering algorithm at a simulated fixed-time controlled intersection that is controlled by a set of signal plans. The used simulation had the following characteristics.

- The signal control strategy in this simulation used a set of six pre-timed signal plans, illustrated in Table 4.10.
- Again, two types of errors are emulated in both scenarios. First, a GPS error $E_{GPS,k}$ for each trajectory k has been applied (see Equation 4.15 in Section 4.3). Second, a proportion of the input data are random trajectories $e_{rate} = 0.2$, used to prove the robustness of developed frequency threshold f_{thr}.
- The simulation emulates an FCD source that supplies two different through movements $m = \{2, 8\}$ of an intersection with very different trajectory volumes. Movement $m = 8$ covers a yearly workday aggregate of approximately 4,000 trajectories, whereas movement $m = 2$ covers only

Signal plan	Time of day [h]	C [secs]	tg_{start} [sec]	tg_{end} [sec]
Plan 1	0 - 4	30	4	16
Plan 2	4 - 7	60	3	33
Plan 3	7 - 10	85	4	51
Plan 4	10 - 15	60	1	30
Plan 5	15 - 19	85	1	47
Plan 6	19 - 24	45	1	22

Table 4.10: Signal plans to evaluate result clustering

approximately 2,000 trajectories. The general demand profile over the time of day is for both movements identical and attached in Appendix C.4, see Figure C.4.11.

Figure 4.22 presents the simple tg_{si} and combined tg green interval estimates for both movements as well as the result of the developed clustering procedure for the first two iterations. Due to less trajectory volumes, estimation results on movement $m = 2$ are fragmented and during some hours estimates are even missing (see Figure 4.22a). In comparison thereto, movement $m = 8$ is able to infer estimates nearly during the complete workday period. In this simulated example, the result clustering and time assignment algorithm should be able to identify large clusters of similar results on movement $m = 8$. The corresponding cluster hours can be used to update the time assignment table (TA) and the process of signal timing calculation should be repeated in order to infer more results on movement $m = 2$.

Column cls_{best} in Figure 4.22a denotes the best found clusters of both movements for all available estimates. For reasons of the algorithm implementation, notation of cls_{best} is slightly different than introduced. The first digit in cls_{best} describes an iteration counter that logs how often the best cluster search has been repeated to cover all hours of estimates. The second digit after separator sign "_" denotes the movement in which the cluster has been found. The third digit expresses the cycle length of the found cluster. The fourth digit is used for internal reasons in the implementation. As expected, the algorithm found in movement $m = 8$ larger clusters than in movement $m = 2$ and proposes to recalculate the signal timing estimates according to a new time assignment $TA_{i=1}$. The colors in 4.22a express sets of hours that belong to the same cluster. A very important consideration at this stage is that the hours $h = \{7, 8, 9\}$ and $h = \{15, 16, 17, 18\}$ belong in time assignment table $TA_{i=1}$ to two different clusters because maximal distance between cluster "4_8_85_1" and "2_8_85_2" is larger than $\delta t_{cluster} = 4$ secs. Figure 4.22b shows the estimation results after the first recalculation of signal timing estimates based on the updated time assignment table $TA_{i=1}$. The successfully clustered hours and therefore updated assignment table allowed the estimation process to operate on an aggregated data set. Thus, even movement $m = 2$ satisfies constrains of data coherence and is able to infer signal timing estimation results, i.e. previous missing hours could be filled up.

A new search process of similar estimation result is started and the time assignment table is updated for a second iteration $TA_{i=2}$. The accepted time tolerance of $\delta t_{cluster} = 4$ secs allows the clustering of hours $h = \{7, 8, 9, 15, 16, 17, 18\}$ that belongs in fact to two different signal plans (plan 3 and 5). The signal timing estimates are again recalculated for the newly updated time assignment table $TA_{i=2}$. As a third update of the time assignment table generates no changes in the cluster structure, method stops. The last result of this example is attached in Appendix C.4.

Movement = 8						Movement = 2								
h	**C**	**tg_{start}**	**tg_{end}**	**$tg_{start,si.}$**	**$tg_{end,si}$**	**h**	**C**	**tg_{start}**	**tg_{end}**	**$tg_{start,si.}$**	**$tg_{end,si.}$**	**cls_{best}**	**TA_{inital}**	**$TA_{i=1}$**
6	60	1	32	1	32	6						1_8_60_1	6	6
7	85	3	46	1	46	7						4_8_85_1	7	7
8	85	4	46	3	46	8						4_8_85_1	8	8
9	85	4	49	3	49	9	85	5	23	5	23	4_8_85_1	9	9
10	60	1	32	59	32	10	60	2	29	1	29	1_8_60_1	10	10
11	60	59	33	58	33	11						1_8_60_1	11	11
12	60	60	32	57	32	12						1_8_60_1	12	12
13	60	1	35	58	35	13	60	1	30	1	30	1_8_60_1	13	13
14	60	2	32	1	32	14	60	59	26	59	26	1_8_60_1	14	14
15	85	85	45	83	45	15						2_8_85_2	15	15
16	85	85	44	84	44	16	85	3	18	3	18	2_8_85_2	16	16
17	85	1	44	83	44	17	85	2	44	2	44	2_8_85_2	17	17
18	85	1	45	83	45	18	85	1	44	1	44	2_8_85_2	18	18
19	45	45	24	42	24	19	45	45	25	44	25	3_8_45_1	19	19
20	45	43	23	42	23	20	45	42	23	42	23	3_8_45_1	20	20
21	45	44	22	43	22	21						3_8_45_1	21	21

(a) First iteration

Movement = 8						Movement = 2								
h	**C**	**tg_{start}**	**tg_{end}**	**$tg_{start,si.}$**	**$tg_{end,si.}$**	**h**	**C**	**tg_{start}**	**tg_{end}**	**$tg_{start,si.}$**	**$tg_{end,si.}$**	**cls_{best}**	**$TA_{i=1}$**	**$TA_{i=2}$**
6	60	60	34	58	34	6	60	1	33	57	33	1_8_60_1	6	6
7	85	3	50	81	53	7	85	4	51	2	51	2_8_85_1	7	7
8	85	3	50	81	53	8	85	4	51	2	51	2_8_85_1	8	8
9	85	3	50	81	53	9	85	4	51	2	51	2_8_85_1	9	9
10	60	60	34	58	34	10	60	1	33	57	33	1_8_60_1	10	10
11	60	60	34	58	34	11	60	1	33	57	33	1_8_60_1	11	11
12	60	60	34	58	34	12	60	1	33	57	33	1_8_60_1	12	12
13	60	60	34	58	34	13	60	1	33	57	33	1_8_60_1	13	13
14	60	60	34	58	34	14	60	1	33	57	33	1_8_60_1	14	14
15	85	85	47	81	49	15	85	85	45	84	45	2_8_85_1	15	15
16	85	85	47	81	49	16	85	85	45	84	45	2_8_85_1	16	16
17	85	85	47	81	49	17	85	85	45	84	45	2_8_85_1	17	17
18	85	85	47	81	49	18	85	85	45	84	45	2_8_85_1	18	18
19	45	44	24	41	25	19	45	45	23	42	23	3_8_45_1	19	19
20	45	44	24	41	25	20	45	45	23	42	23	3_8_45_1	20	20
21	45	44	24	41	25	21	45	45	23	42	23	3_8_45_1	21	21

(b) Second iteration

Figure 4.22: First and second iteration of the result clustering

In summary it can be said that with the help of this simulation experiment the developed result clustering and time assignment algorithm has been successfully tested and the proposed algorithm showed the desired functionality. Fragmented estimation results of movement $m = 2$ benefit from parameter patterns found in movement $m = 8$. This experimental study reveals also an important side effect. Due to the recalculation of the signal timing estimates based on the updated time assignment table, results can come in each iteration closer till they remain under the specified time tolerance of $\delta t_{cluster} = 4$ and are therefore also merged. Thus, this method brings advantages and disadvantages. On the one hand, the concept of result clustering and time assignment is

a fundamental process that allows to derive estimates on movements with less trajectory volumes. On the other hand, the tolerance which is needed for the clustering process $\delta t_{cluster}$ allows likewise the grouping of hours which are in fact controlled by slightly different signal plans. As this thesis has the objective to prove and prevent overlapping green intervals of conflicting movements, estimates on movements with less trajectory volumes are mandatory and the described disadvantage needs to be accepted.

4.7 Verification of Conflicting Green Intervals

4.7.1 Intergreen Time Calculation

The last processing stage of the developed signal timing estimation process covers a verification procedure that proves potential contradiction between green intervals of conflicting movements. Because the estimation process has on an intersection with partially missing results no reliable evidence whether a turning movement is in reality controlled by the permissive or protected mode, this verification procedure checks only conflicts between movements that need to be controlled by different phases (protected mode). As the green intervals of conflicting movements are typically separated by a certain time interval, i.e. the intergreen time, this subsection summarizes briefly the German approach of the intergreen time calculation. Furthermore, needed data requirements to calculate intergreen times at real-world intersections based on the Open Street Map graph model are introduced [33].

The intergreen time calculation specifications used in this thesis refers to the German guideline "RiLSA" [28]. As the signal timing estimation process analyzes trajectories of motorized vehicles, ongoing explanations consider only two intergreen time calculation scenarios that are related to the motorized traffic. The first scenario describes the case in which the through movement of an intersection is clearing, the second scenario corresponds to the case in which a turning movement is clearing the intersection area. The basic calculation rules for both scenarios are equivalent but used constants differ slightly. For the sake of brevity, this paragraph introduces the calculation method for the first scenario and provides for both scenarios necessary constants (see Table 4.11). The used notations refer to the research from Rezko et al. [69] who summarized very well the German approach of intergreen time calculation.

According to Rezko et al., variable to denotes the overrun time which describes the duration between the end of the green interval and the point of time at which the last vehicle of the ending green interval passes the stop bar. Clearing time, tc expresses the duration which is needed by the last vehicle of the ending green interval to clear the distance to the conflict point dc including its own vehicle length l_{veh}. The entrance time, te describes therefore the duration for the first vehicle of the upcoming green interval to traverse the

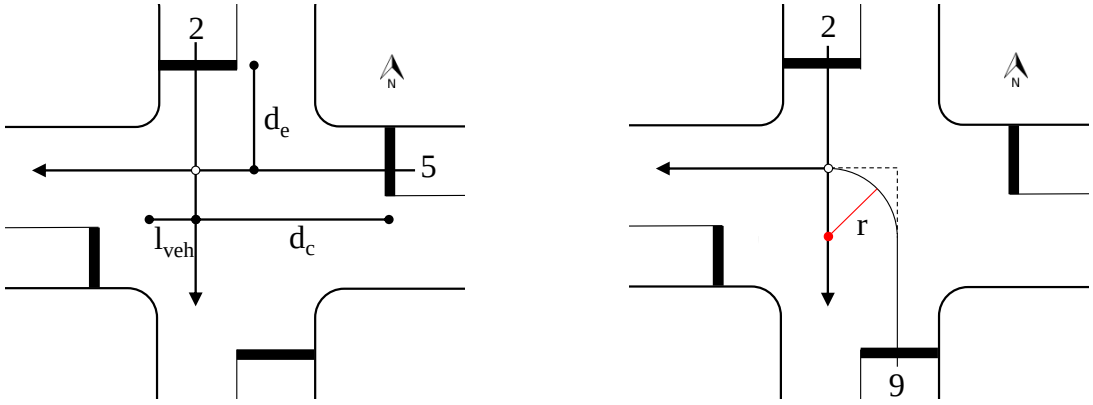


Figure 4.23: $m = 5$ is clearing and $m = 2$ is entering (a), $m = 9$ is clearing and $m = 2$ is entering (b)

entrance distance de. Figure 4.23a illustrates the mentioned parameters that are used in Equation 4.72 to calculate the intergreen time.

$$ti = to + tc - te \tag{4.72}$$

$$tc = \frac{dc + l_{veh}}{vc} \tag{4.73}$$

$$te = \frac{3.6 \cdot de}{ve} \tag{4.74}$$

The introduced calculation concept makes clear that stop bars as well as the clearing and entrance distances to the conflict point are required information to calculate the intergreen time at real-world intersections. Referring to Subsection 4.1.1, the stop bar information is already available for each movement and is supplied manually by the user. Thus, to determine the clearing and entrance distances, the method of this thesis proposed to make use of the intersection model, described in Subsection 4.1.1. This model stores each movement as a sequence of edges that are obtained from Open Street Map [33]. The conflict point represents thus the spatial intersection point of two conflicting movements and can be calculated by linear algebra or with the help of a geographic information system [67] that is used to store the road graph and intersection model.

Table 4.11: Constants for intergreen time calculation [28]

Units	Through movement clears	Turning movement clears
to [secs]	3	2
ve [km/h]	40	40
vc [m/s]	10	6
l_{veh} [m]	6	6

As the graph model is obtained from the Open Street Map database, the road geometries are only a very generalized version of real movement layout. Which means, derived intergreen times may deviate from ground truth values. The generalized road graph of Open Street Map affects especially the intergreen times of turning movements, illustrated in Figure 4.23b. The perpendicular dashed lines in Figure 4.23b express a simplified movement geometry, obtained from Open Street Map. The depicted examples show that a turning vehicle needs to traverse in the Open Steet Map graph an additional distance s_{er}. A theoretical consideration assesses the potential of this additional distance and estimates corresponding time errors in the clearing tc_e and the entering time te_e when using such a simplified road graph for the intergreen time calculating.

For a radius of $r \approx 20$ m, Equation 4.75 returns a distance error of $s_{er} \approx 8.6$ m, which causes in case of Figure 4.23b for movement $m = 9$ a potential clearing time error of $tc_e \approx 1.4$ secs and an entering time error of $te_e \approx 0.8$ secs. These error potentials need to be tolerated in order to incorporate the intergreen time estimates in the ongoing green interval verification procedure. However, absolute error values are relatively small when comparing these against the effect of trajectory interpolation.

$$s_{er} = 2r - 1/2\pi r \tag{4.75}$$

$$tc_e = \frac{s_{er}}{vc} \tag{4.76}$$

$$te_e = \frac{3.6 \cdot s_{er}}{ve} \tag{4.77}$$

The intergreen time of conflicting movements are stored in a square matrix A that holds in $a_{i,j}$ the required intergreen time for the scenario that a vehicle from movement i is clearing and another vehicle from movement j is entering the intersection area. The number of movements fits to a typical 4-arm intersection.

$$A \in \mathbb{N}^{n \times n}, n = |m|, m = \{1, \ldots, 12\} \tag{4.78}$$

4.7.2 Verification and Adjustment Procedure

The previous subsection has worked out the data requirements for the intergreen time calculation at real-world intersections. The objective of this subsection is to describe an approach that checks the green intervals of conflicting movements in order to prevent green interval conflicts in consistency with the previously calculated intergreen times. An overlapping of green intervals can appear for two reasons and needs to be prevented by adjusting the estimated green intervals.

1. **Inaccurate green intervals:** Simulative studies in Subsection 4.5.1 quantified the estimation result variability for different traffic signal control

scenarios. These findings show that the method inherent result variability can cause overlapping green intervals.

2. **Actuated traffic signal control:** Actuation causes typically dynamic changes in the phase sequence and green time durations. As the basic concept of the signal timing estimation procedure aggregates trajectory crossing events and stop seconds by folding a long historical time period, conflicting movements may thus contain cycle seconds in which both movements claim to have green. Such a situation is depicted in Figure 4.24, in which the green interval of movement $m = 5$ has a conflict with movement $m = 2$.

To adjust overlapping green intervals, the method proposes to compare a particular green interval tg_i at movement i with its conflicting movements j. This pairwise comparing process is based on the histograms $h(b)_i$ and $h(b)_j$ that are generated in the simple or combined estimation procedure (see Section 4.2). Within this process, so called center points $cp_{i,j}$ and/or intersection points $ip_{i,j}$ between green intervals of conflicting movements i, j are calculated. Both point types are used as orientation time points to fit the intergreen times $a_{i,j}$ and $a_{j,i}$ into the cycle time by symmetrically adjusting adjacent green intervals of movement i and j. The objective of this procedure is to generate green intervals that are consistent with previously calcualted intergreen times. Moreover, if the checked green intervals need to be heavily adjusted in order to achieve the desired consistency, green intervals will be assumed to be highly traffic actuated.

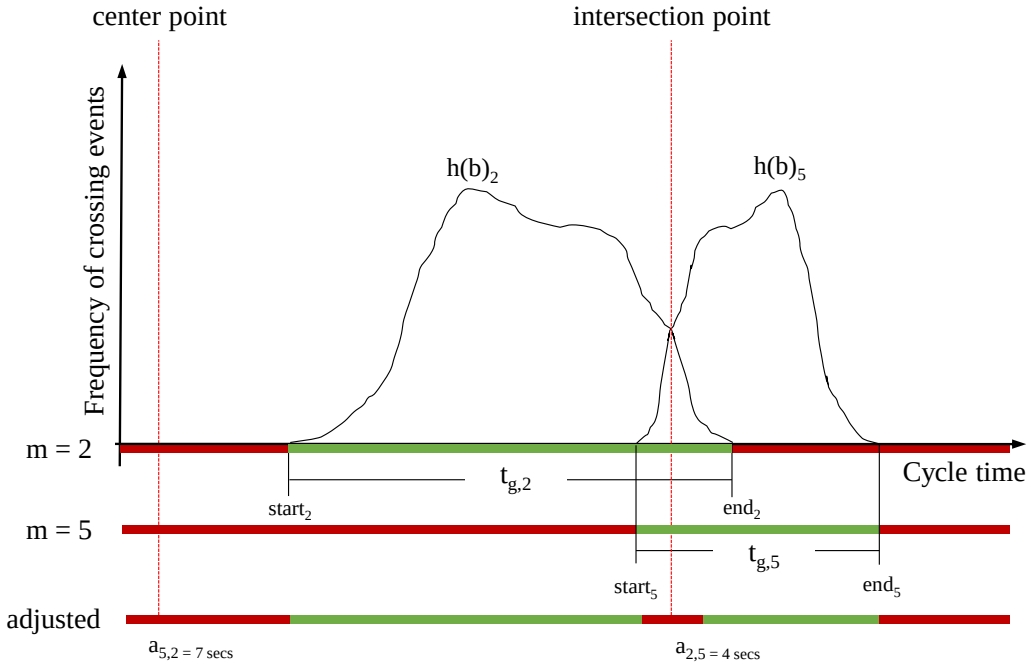


Figure 4.24: Concept of center and intersection points to adjust green intervals

Figure 4.24 illustrates an example in which the green intervals of two conflicting movements $i = 2$ and $j = 5$ are shown over the cycle time. The ending green interval of movement $i = 2$ overlaps the beginning green interval of movement $i = 5$. Which means that both histograms need to have an intersection point that lies in between the end of tg_2 and the start of tg_5. This intersection point is indicated by a red vertical dashed line in Figure 4.24. In contrast thereto, between the ending interval of tg_5 and the starting green interval of tg_2 is a longer red period, whereas the center of this red period is the previously mentioned center point. The colored bars below the abscissa in Figure 4.24 depict the original green interval for each movement as well as the resulting green intervals that match the required intergreen times, labeled with the word "adjusted".

Alternatively to the described approach, the green start of movement $j = 5$ could be also used as an orientation point for the green interval adjustment process. Such a rule can be adverse, if the green start of movement $j = 5$ or the end of movement $i = 2$ are intensively traffic actuated. In such a situation it would be more reasonable, to select a time point that lies in between the traffic actuated parts of both adjacent green intervals, which is reflected by the use of the previously defined intersection point.

The principle calculation of center or intersection points for two conflicting green intervals requires the precondition that both green intervals are not allowed to be nested, expressed by Equation 4.79. The green interval tg_i and tg_j are thereby interpreted as a set of seconds. If the mentioned precondition is violated, the combination of both green intervals tg_i and tg_j is labeled as a problematic conflict that can not be resolved and needs a manually inspection. Furthermore, the adjustment process defines also a postcondition that requires that an adjusted green interval $\bar{t}g_i$ of movement i needs to satisfy a minimum green interval of $\delta min_t > 7$ secs. If the postcondition is violated, the result can be used for further applications, but it should be labeled in order to allow the application a result exclusion.

Precondition for green interval adjustment: (4.79)

$$tg_i \not\subset tg_j \wedge tg_j \not\subset tg_i$$

Postcondition for adjusted green interval: (4.80)

$$\bar{t}g_i > \delta min_t$$

The variable tg_i denotes the unadjusted green interval for a movement, whereas a green interval corresponds to its binary state vector $state_i$ and has a starting and ending second, denoted with $[start, end]$. According to the previous explanations, the center point calculation $cp_{i,j}$ between the ending green interval of movement i and the starting green interval of movement j is only allowed, if both green intervals are separated by a red interval, expressed by Equation 4.82. The expression $[0 \ldots 0]$ denotes a slice of seconds in the state

vector $state_i$ in which all seconds belong to the signal state red.

$$tg_i = [start, end] \tag{4.81}$$

$$cp_{i,j} \quad | \; state_i\,[end_i + 1, \ldots, start_j - 1] = [0 \ldots 0] \tag{4.82}$$
$$\wedge \; state_j\,[end_i + 1, \ldots, start_j - 1] = [0 \ldots 0]$$

$$ip_{i,j} \quad | \; state_i\,[start_j \ldots, end_i] = [1 \ldots 1] \tag{4.83}$$
$$\wedge \; state_i\,[start_j \ldots, end_i] = [1 \ldots 1]$$

If the condition in Equation 4.82 is violated, both green interval are overlapping and need to have an intersection point $ip_{i,j}$. The calculation of the center point $cp_{i,j}$ between movement i and j is defined by Equations 4.85 and 4.86.

$$cp_{i,j}(tg_i, tg_j) = \begin{cases} eq.\,4.85, & \text{if } start_j < end_i \\ eq.\,4.86, & \text{otherwise} \end{cases} \tag{4.84}$$

$$= \lceil \big(end_i + (start_j + C - end_i)/2\big) \rceil \mod C \tag{4.85}$$

$$= \lceil \big(end_i + (start_j - end_i)/2\big) \rceil \mod C \tag{4.86}$$

In contrast to the center point $cp_{i,j}$, an intersections point $ip_{i,j}$ is calculated by a second-wise comparison of histogram counts per bin b from $h(b)_i$ and $h(b)_j$, obtained by the simple or combined green interval estimation methodology. Thus, the time point of intersection $ip_{i,j}$ between the ending green of movement i and the starting green of movement j is calculated with function $ip_{i,j}\big(h(b)_i, h(b)_j\big)$. An intersection point is searched within a bin slice, denoted by $b[start_j \ldots end_i]$. The intersection point $ip_{i,j}$ is found, if $h(b)_j$ exceeds or is equal to $h(b)_i$. Volatile histograms could generate multiple intersection points. Thus, methods are needed to obtain only one particular intersection point. This could be achieved for instance by clustering and averaging found intersection points within the bin slice $b[start_j \ldots end_i]$.

$$ip_{i,j}\big(h(b)_i, h(b)_j\big) = \text{function that returns one intersection point} \tag{4.87}$$

$$\text{with} \quad \big(h(b)_i \quad | \quad b[start_j \ldots end_i]\big) \wedge \big(h(b)_j \quad | \quad b[start_j \ldots end_i]\big) \tag{4.88}$$

The calculated center and intersection points are used to adjust the green intervals with respect to the estimated intergreen times $a_{i,j}$ and $a_{j,i}$. Equation 4.90 calculates as an example the adjusted green interval $\hat{tg}_i$ for the situation that tg_i is intersecting on both interval sides with its conflicting green interval tg_j.

$$\hat{tg}_i = [start, end] \tag{4.89}$$

$$start_i = (ip_{j,i} + \lceil a_{j,i}/2 \rceil) \mod C \tag{4.90}$$

$$end_i = (ip_{i,j} - \lceil a_{i,j}/2 \rceil) \mod C \tag{4.91}$$

As the verification method checks an individual green interval tg_i for a particular movement i against all its conflicting movements j, multiple adjusted

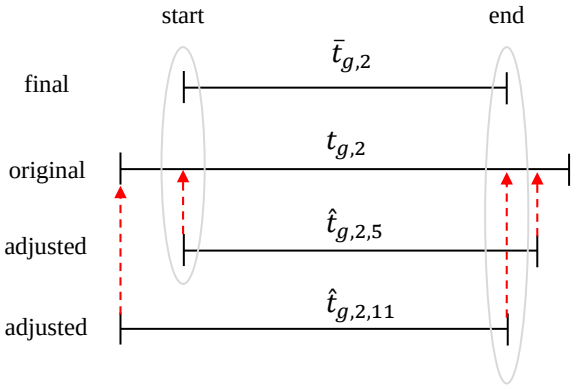


Figure 4.25: Final green interval selection procedure

green intervals for movement i are calculated. Each conflicting movement j may have different intergreen time requirements and thus, a final procedure is needed to figure out, which conflicting movement j has caused the most severe changes on the start and the end of the original green interval tg_i. This procedure compares tg_i against all available adjusted intervals $\hat{t}g_{i,j}$ by calculating the absolute deviation for the start and the end of green before and after the adjustment procedure according to Equation 4.92. The variable $\bar{t}g_i$ represents thus the final estimated green interval with its start and end for movement i.

$$\bar{t}g_i = [start, end] \tag{4.92}$$

$$start_i = \forall j \; arg\,max_{\hat{t}g_{start,i,j}} \big((\hat{t}g_{start,i,j} - tg_{start,i}) \mod C\big) \tag{4.93}$$

$$end_i = \forall j \; arg\,max_{\hat{t}g_{end,i,j}} \big((tg_{end,i} - \hat{t}g_{end,i,j}) \mod C\big) \tag{4.94}$$

The described process is finally depicted in Figure 4.25, in which tg_2 represents the original green interval, deduced with the simple or combined green interval estimation methodology. The intervals $\hat{t}g_{2,5}$ and $\hat{t}g_{2,11}$ represent two adjusted green interval solutions for movement $i = 2$, whereas the solution $\hat{t}g_{2,5}$ requires the largest adjustment of the green start and solution $\hat{t}g_{2,11}$ requires the largest adjustment on the green end of tg_2 in order to satisfy the intergreen times.

4.8 Composed Estimation Process and Result Classes

All previous sections in this chapter explained consecutively the necessary tasks to infer signal timing parameters on real-world intersections by exploiting low frequency trajectory volumes. For reasons of comprehensibility and to understand the interdependencies between different processing tasks, this section closes the methodology development with a composed flowchart of all required processing steps, whereas the functionally of each step is summarized briefly.

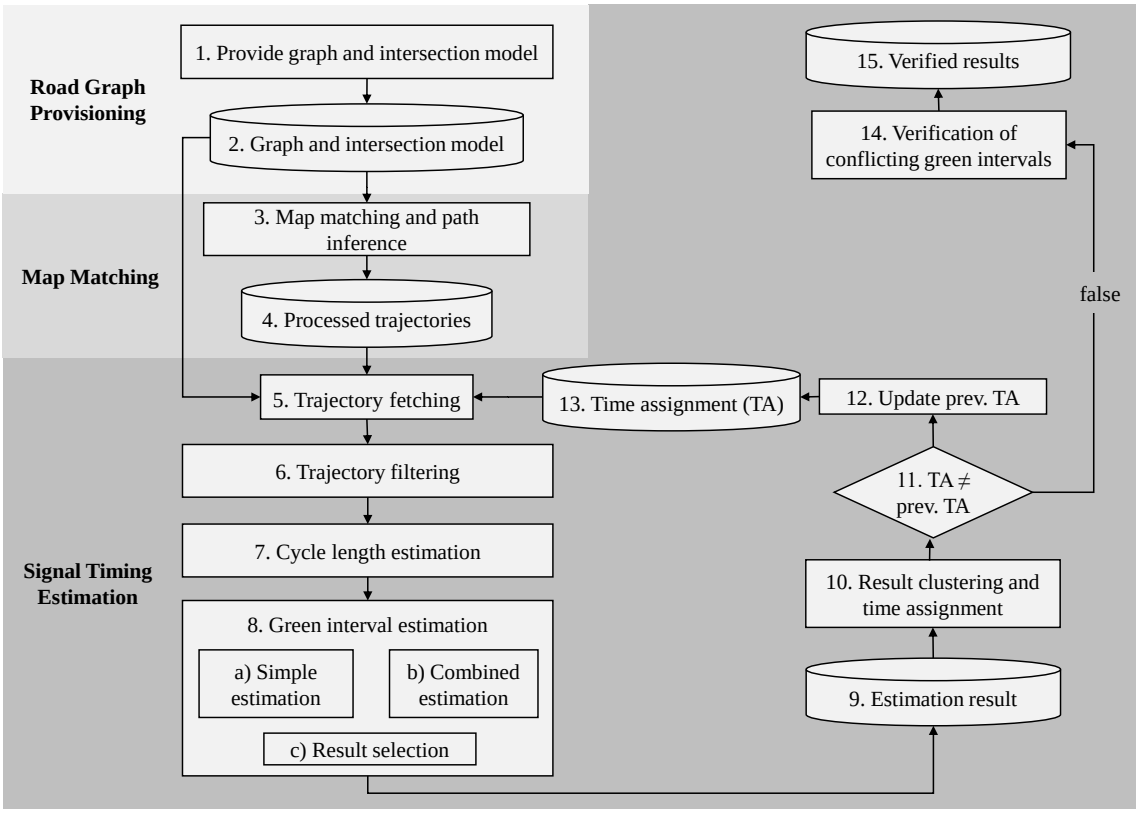


Figure 4.26: Composed signal parameter estimation process

- **Graph and intersection model:** The methodology starts by the provisioning of a graph model and the manual definition of the intersection model (1-2).
- **Map matching and path inference:** The provided graph model is thereby used to infer the most likely driven path of each vehicle and to reference trajectory data points logically with the graph model (3). All matched trajectories are stored and in a spatial database (4).
- **Trajectory fetching:** The fetching process (5) has the task to query trajectories from the spatial database (4) with respect to the time assignment table (12) for a certain intersection of interest. The relationship between trajectories and each intersection is managed with the help of the intersection model (2). The time assignment table stores hours in which traffic signal control is assumed to have constant signal parameters. In the first initial estimation iteration, the process assumes constant parameters for every hour of the analyzed yearly workday period.
- **Trajectory filtering:** A filtering process (6) tries to reduce the number of erroneous trajectories on through movements caused by position inaccuracy and trajectory interpolation.

- **Cycle length estimation:** The cycle length estimation procedure estimates for each hour of the workday a cycle length which allows to reject an assumed uniform distribution of crossing events at least with a level of significance of $\alpha = 0.005$. If multiple cycle length candidates are in the choice set, the most significant result is used for further data analysis (7).

- **Green interval estimation:** With the knowledge of the cycle length, the green interval estimation procedure tries to infer separately a contiguous green interval for every individual movement of an intersection. The estimation process distinguishes two methods, simple (8 a) and combined estimation (8 b). If no combined estimation result can be inferred, the procedure uses alternatively for all ongoing processing steps a simple estimation result (8 c).

- **Result clustering and time assignment:** The derived estimation results for each movement are checked for similarities and conditionally clustered. An iterative cluster selection process tries to minimize the total number of signal plans at an intersection and creates a new time assignment table (9-13). A newly derived time assignment table is used to provide the estimation process aggregated time intervals in which signal control parameters seem to be very similar and a new set of signal parameters is estimated. If the clustering process is not able to generate a new time assignment table, results are forwarded to the next processing step (14).

- **Verification of conflicting green intervals:** In a last processing step, the green interval estimates of conflicting movements are verified against each other. The manually defined intersection model is exploited to derive intergreen time between conflicting movements that are required within the green interval verification procedure. Results are finally stored for calibration and validation processes (14-15).

The composed estimation procedure is capable to return signal timing estimates with different properties that can be expressed by various result classes. The achieved class for movement $m = i$ depends on the used green interval estimation method (simple or combined) as well as on the capability that a result could be verified at least by one conflicting movement $m = j$. Later in later in Chapter 5, these classes are introduced to allow statements of the estimation quality by considering the achieved result class.

1. **Class 1:** The first result class describes green interval estimates derived with the help of the simple estimation methodology for a single movement $m = i$. The result describes a contiguous green interval determined by considering only crossing events. Due to missing results on any conflicting movement $m = j$ no result verification is possible at all.

Table 4.12: Overview of possible result classes

$class$	Intergreen verified	Method at $m = i$	Method at $m = j$
1	false	simple	-
2	false	combined	-
3	true	simple	simple
4	true	simple	combined
5	true	combined	simple
6	true	combined	combined

2. **Class 2:** The second result class describes green interval estimates derived with the help of the combined estimation methodology for a single movement $m = i$. The result describes a contiguous green interval determined by considering crossing events and complementary stop seconds. Due to missing results on any conflicting movement $m = j$, no additional result verification is possible.

3. **Class 3:** The third result class describes the case in which a simple estimation result of movement $m = i$ can be at least verified by a single result of the same result class on a conflicting movement j.

4. **Class 4:** The fourth result class describes the case in which a simple estimation result of movement $m = i$ can be at least verified by a single combined estimation result on a conflicting movement $m = j$.

5. **Class 5:** The fifth result class describes the case in which a combined estimation result of movement $m = i$ can be at least verified by a single simple estimation result on a conflicting movement $m = j$.

6. **Class 6:** The sixth result class considers the best available result class. A combined estimate on movement $m = i$ can be successfully verified against at least one conflicting movement that had the same result class.

CHAPTER 5

Evaluation

5.1 Calibration

5.1.1 Objectives

This section is dedicated to the calibration procedure of the developed signal timing estimation methodology. Beforehand going into the details, terms and objectives of calibration and validation are introduced briefly referring to the definitions of German Road and Transportation Research Association [27].

Definition 1 (Calibration)**.**
Calibration is the systematic adjustment process of model parameters with the objective to replicate a ground truth situation in such a manner that the deviations between the model and the ground truth are within an acceptable tolerance.

Definition 2 (Validation)**.**
Validation is the process to provide evidence that the calibrated model is capable to replicate another ground truth situation within a specified tolerance when applying the calibrated model on another different dataset.

Several simulation studies have been carried out in this thesis that allowed to demonstrate the basic functionality and to make findings of the result variability with respect to different sample sizes and signal control strategies. Furthermore, the studies emphasizes the influence of movement's degree of saturation, when green intervals are calculated with the help of the combined estimation methodology. Up to this step, all established studies are based on the very same setup of calibration parameters. The determination of this model parameters followed thereby no strict scheme, which is argued by the model complexity. Nevertheless, the applied parameters allowed for very different scenarios the replication of the ground truth start and end of green in simulation within an median deviation of approximately 5 secs (see Section 4.5.1 and 4.5.2). Thus, it can be stated that the used parameter combination is

Table 5.1: Overview of potential calibration parameters

Parameter	Value [units]	Functionality	Paragraph
$w(s)$	function()	Weighting function that weights a stopping event with respect to position in tailback	4.2.2
n_{lvl}	15 [-]	Number of levels to scale frequency threshold, used to classify a cycle second to a green or a red interval	4.5.1
$iteration_{nr}$	2 [-]	Allowed number of iterations to satisfy constrains of data coherence	4.5.1
$\delta const_t$	2 [secs]	Required interval to allow compensation of peaks and gaps in second-wise state vector	4.5.1
δmin_t	7 [secs]	Minimum red and green intervals to satisfy constraints of data coherence.	4.5.1
w_{red}	$1/4$ [-]	Weighting factor used to level crossing events and stop seconds with the objective to compensate effects of interpolation error	4.5.2
$\delta t_{cluster}$	4 [secs]	Time tolerance for green start and end to allow clustering of similar green intervals	4.6

least validated for the tested simulation scenarios. Table 5.1 presents a list of all calibration parameters including the set values. Furthermore, Table 5.1 summarizes briefly the basic functionality of each parameter within the estimation procedure. One should remark that a systematical analysis of all available calibration parameters for a set of real-world intersections requires a highly efficient implementation of the developed process, which has not been the main objective of this research.

Instead, this thesis has the objective to work out the basics of an estimation procedure that is capable to infer signal parametrization considering sparse trajectory volumes, sampled with low frequencies. Thus, the calibration procedure of this section focuses only on one specific parameter that is relevant for the green interval estimation and has already proved to have a high importance. This parameter is w_{red}, highlighted in Table 5.1. The calibration procedure is carried out for two real-world intersections in Munich (Germany) by considering a set of five different weights, $w_{red} = \{1, 1/2, 1/4, 1/8, 0\}$. Most

Table 5.2: Fact sheet: Ingolstädter Straße / Frankfurter Ring

Intersection id:	13
Major road:	Ingolstädter Straße
Signal control type:	actuated
Signal phase structure:[1]	Ph1=1,2,3,7,8,9 Ph2=1,10,11,12 Ph3=4,5,6,7
Public transport:	Bus
Daily traffic volume[2]	45/40
Intersection layout:	4 arms
Ground truth period:	07-11/07/2014
Specific features:	partially grade-separated

[1] movement numeration schema (see Section 3.2)
[2] units in veh/day for major/minor road

Table 5.3: Fact sheet: Ingolstädter Straße / Neuherbergstraße

Intersection id:	2
Major road:	Ingolstädter Straße
Signal control type:	actuated
Signal phase structure:[1]	Ph1=1,2,3,7,8,9 Ph2=4,5,6,10,11,12
Public transport:	Bus
Daily traffic volume[2]	40/22
Intersection layout:	4 arms
Ground truth period:	07-11/07/2014
Specific features:	-

[1] movement numeration schema (see Section 3.2)
[2] units in veh/day for major/minor road 3.2)

important proprieties of these two intersections are summarized in Table 5.2 and Table 5.3.

The first intersection in the calibration dataset is located at Ingolstädter Straße / Frankfurter Ring and has the special road design that movements $m = \{5, 11\}$ are partially grade-separated which means traffic is designated to use a ramp which allows to pass the intersections' area. Nevertheless, vehicles may also cross the intersection on the ground-level to reach local shopping facilities. The map matching and path inference algorithm is not able to distinguish reliably whether a vehicle has passed this intersection on the ground-level or via the ramp. This means not all crossing events on these two movements were regulated by the signal control and thus, movements $m = \{5, 11\}$ are excluded in ongoing data analysis.

The second intersection in the calibration dataset is located at Ingolstädter Straße / Neuherbergstraße which has likewise to first intersection four arms, whereas the eastern arm enters a military facility and thus movements that are starting or ending in this arm are excluded in data analysis because traffic volumes into and from this area are negligible.

The trajectory dataset used in the calibration and validation procedure covers a period of one year (03/2014-03/2015). Most important data characteristics already have been introduced in Section 3.1. The available trajectories volumes for both intersections are documented in Appendix D.1.2 and D.2.2. The ground truth signal states in the calibration and validation procedure have been obtained from the City of Munich.

5.1.2 Cycle Length Estimation

Because the calibration procedure involves a model parameter of the green interval estimation, this section is first dedicated to the results of the cycle length estimation for both traffic actuated intersections, mandatory for any green interval estimation.

The developed cycle length estimation methodology of Section 4.4 has been used to infer for the mentioned time interval the typical cycle length for the workday period with an hourly time resolution by testing potential cycle length candidates in a range from 30 to 120 secs. The method has thereby been repeated for two of the most common backward calculation methods in Germany (RRV and RRV 1.1., see Appendix F).

Table 5.4 presents the results of this procedure, whereas only the backward calculation method "RRV" allowed the inference of the cycle length $C = 70$ secs at both intersections. This can be explained as follows. If the local time switches from summertime (UTC+1) to wintertime (UTC+2), the backward

Table 5.4: Estimated (C) and ground truth cycle length (C_{true}) at Ingolstädter Straße / Frankfurter Ring

h UTC	m	C [sec]	C_{true} [sec]	A^2 [-]	p_{value} [-]	total vol. [veh/h]
00:00:00	2	-	70	-	-	83
01:00:00	8	-	70	-	-	92
02:00:00	2	-	70	-	-	81
03:00:00	2	-	70	-	-	224
04:00:00	2	90	90	18.06	5.22e-06	511
05:00:00	2	90	90	25.68	3.75e-06	504
06:00:00	2	90	90	60.21	3.95e-06	503
07:00:00	2	90	90	48.43	3.85e-06	421
08:00:00	2	90	90	40.53	4.08e-06	386
09:00:00	8	90	90	8.94	3.88e-05	510
10:00:00	2	90	90	7.82	1.41e-04	443
11:00:00	8	90	90	78.54	2.71e-06	477
12:00:00	8	90	90	86.54	2.55e-06	494
13:00:00	2	90	90	91.10	3.59e-06	469
14:00:00	2	90	90	98.87	3.49e-06	469
15:00:00	8	90	90	83.86	2.40e-06	495
16:00:00	2	90	90	136.46	2.34e-06	636
17:00:00	2	90	90	134.57	2.09e-06	642
18:00:00	2	90	90	59.40	3.28e-06	401
19:00:00	8	70	70	24.92	3.92e-06	300
20:00:00	8	70	70	35.80	5.22e-06	262
21:00:00	8	70	70	41.38	5.77e-06	228
22:00:00	8	-	70	-	-	158
23:00:00	8	-	70	-	-	107

Table 5.5: Estimated (C) and ground truth cycle length (C_{true}) at Ingolstädter Straße / Neuherbergstraße

h UTC	m	C [sec]	C_{true} [sec]	A^2 [-]	p_{value} [-]	total vol. [veh/h]
00:00:00	2	-	70	-	-	66
01:00:00	2	-	70	-	-	75
02:00:00	2	-	70	-	-	105
03:00:00	2	-	70	-	-	274
04:00:00	2	90	90	26.27	4.58e-06	982
05:00:00	2	90	90	64.10	2.14e-06	1269
06:00:00	2	90	90	72.87	1.90e-06	978
07:00:00	2	90	90	51.06	2.48e-06	751
08:00:00	2	90	90	39.76	2.37e-06	697
09:00:00	8	90	90	30.80	3.66e-06	635
10:00:00	2	90	90	10.13	7.71e-06	631
11:00:00	8	90	90	66.77	1.95e-06	633
12:00:00	8	90	90	78.30	1.86e-06	647
13:00:00	8	90	90	68.29	1.98e-06	660
14:00:00	8	90	90	61.35	1.91e-06	773
15:00:00	8	90	90	68.46	1.68e-06	863
16:00:00	8	90	90	72.20	1.53e-06	859
17:00:00	8	90	90	87.67	1.52e-06	660
18:00:00	8	90	90	45.77	2.23e-06	465
19:00:00	1	90	90	5.63	1.44e-03	271
20:00:00	2	70	70	20.08	5.88e-06	186
21:00:00	8	70	70	24.83	5.50e-06	158
22:00:00	2	-	70	-	-	157
23:00:00	2	-	70	-	-	90

calculation second in the "RRV 1.1." method [56] realizes a time leap of 3,600 sec, whereas the backward calculation second in "RRV" is continuously running steadily. If the modulo based time folding approach assumes this time leap which never happens in reality, the time folding of cycles would be not congruent for a cycle length of 70 secs, because 3,600 secs is not a multiple of 70 secs. As described in Subsection 4.4 the cycle length is always concluded from the movement m that has the largest number of trajectories $totalvol$ for a given hour h. The variable A^2 and p_{value} denotes the used test statistic and the observed level of significance, described in Subsection 4.4. The developed approach seems to perform very well because all estimated cycle lengths C match perfectly with the ground truth values C_{true}. Only in the early morning and late evening, trajectory volumes are not sufficient enough to derive a cycle length estimation with the preassigned significance level of $\alpha = 0.005$. These very good results can mainly be explained by the fact that both analyzed intersections are switching their signal plans to the full hour.

5.1.3 Calibration Procedure

With the knowledge of intersections' cycle lengths, the green interval estimation methodology of Subsection 4.5.1 is applied to estimate hourly for the workday period green intervals for each movement by iterating over different weights $w_{red} = \{1, 1/2, 1/4, 1/8, 0\}$. The parameter $w_{red} = 1$ means thereby that measured stop seconds are used unweighted, whereas $w_{red} = 0$ excludes

Table 5.6: Calibration results for intersection Ingolstädter Straße / Frankfurter Ring, id=13

					QKZ class			
w_{red}	$mn(QKZ_1)$	$mn(QKZ_2)$	$sd(QKZ_1)$	$sd(QKZ_2)$	A	B	C	$\sum$
1	0.52	0.37	0.26	0.25	160	83	159	402
1/2	0.60	0.30	0.23	0.19	192	118	107	417
1/4	0.63	0.30	0.21	0.18	217	98	78	393
1/8	0.69	0.28	0.21	0.18	255	74	64	393
0	0.76	0.29	0.20	0.18	293	32	68	393

all measured stop seconds from the green interval estimation process. All exploited trajectories needed to match the sampling interval of $t_d \leq 15$ sec and trajectories on through movements $m = \{2, 5, 8, 11\}$ have been filtered with the help of the Random Forest classification algorithm to reduce the effect of trajectory interpolation (see Section 4.3).

As the trajectory dataset covers an one year workday time period from 03/2014-03/2015, the ground truth traffic signal states for a systemic comparison have been obtained for both intersections over the time period from 07/07-11/07/2014 (Mon. to Fri.). This time period is free of any vacations or holidays. The ground truth signal states have been provided for each movement as a second-by-second state sequence, whereas Equation 4.64 of Subsection 4.5.2 has been applied to estimate for each ground truth second its corresponding estimated signal state. Finally, estimated and ground truth state sequences have been converted and stored as sets of green intervals in which the signal state amber in the reference data has been rated again as green, because vehicles in Germany are legally allowed to cross the stop bar in this time interval.

The estimated and the ground truth green interval sets have been analyzed according to the QKZ measures, introduced in Section 3.3. The comparison between ground truth and estimated green intervals considers logically only hours of the typical weekday period for which the estimation methodology could infer green intervals. Due to the enormous amount of representable information for each movement and hour, the author decided to show results stepwise, starting with a tabularly overview of the QKZ measures that allow to identify the best weighting parameter w_{red}. Thus, Table 5.6 and Table 5.7 present a global view on the achieved green interval estimation quality by summarizing the arithmetic mean mn and the standard deviation sd of the QKZ measures for both intersections over all movements and hours for which estimates have been available. Both tables show additionally, how many estimates have been allocated to a particular QKZ class. The general calibration objective was to get most estimates within QKZ class A while preserving the total number of available estimates, expressed by the tables' last column $\sum$.

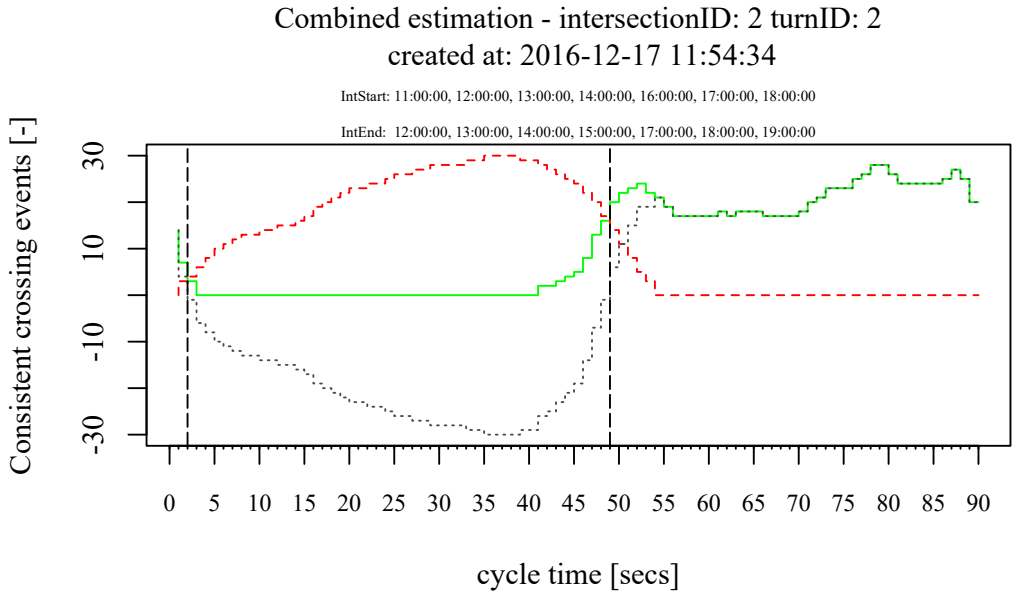
Table 5.7: Calibration results for intersection Ingolstädter Straße / Neuherbergstraße, id=2

					QKZ class			
w_{red}	$mn(QKZ_1)$	$mn(QKZ_2)$	$sd(QKZ_1)$	$sd(QKZ_2)$	A	B	C	$\sum$
1	0.80	0.09	0.13	0.04	212	8	0	220
1/2	0.85	0.08	0.12	0.04	212	8	0	220
1/4	0.89	0.08	0.11	0.03	212	8	0	220
1/8	0.92	0.08	0.09	0.03	212	4	0	216
0	0.92	0.07	0.10	0.04	212	4	0	216

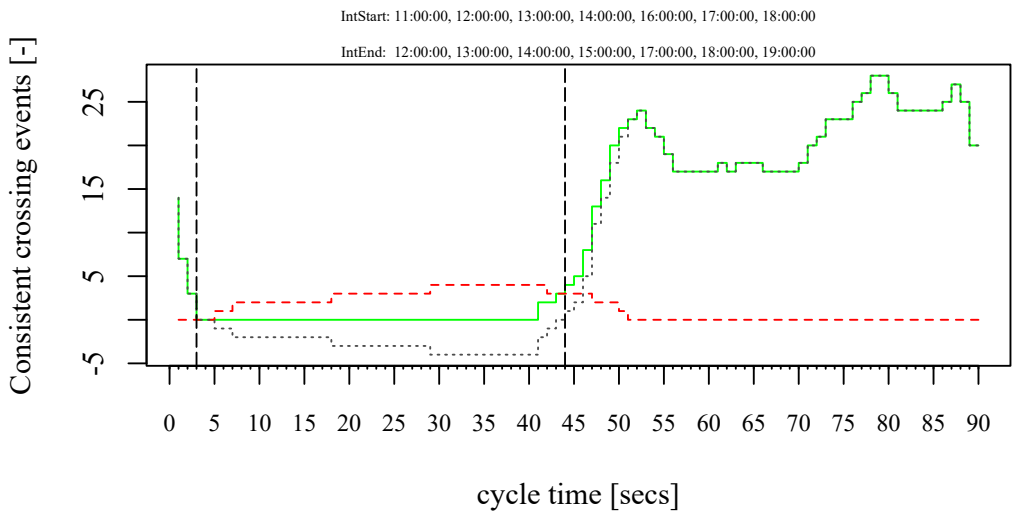
The presented information allow thus the intuitive selection of the weight w_{red} that led to the best estimation quality on both intersections. It can be clearly figured out that $w_{red} = 1/4$, used in previous simulation studies, is not the optimal parameter for these two real-world intersections. Instead, $w_{red} = 0$ allows an substantial improvement of QKZ_1 without impairing QKZ_2 or loosing lots of estimation results at all. Nevertheless, varied weights led to slightly different numbers of available estimation results, summarized by the tables' last column $\sum$. This effect has been analyzed and can be traced back to the fact that changed weights can cause different estimation results for each hour per movement and thus result clustering and time assignment algorithm (see Section 4.6) produces different groups of similar estimation results. If the cluster size covers only one or two hours, minor movements are not able to reveal green interval estimates and thus, the total number of available estimates may differ.

According to the results of Table 5.6 and Table 5.7, the author decided to use $w_{red} = 0$ for the further result analysis. Before going into details, a short excursus has the objective to figure out, why $w_{red} = 0$ allows an improvement of the estimation quality. As the reason needs to be related to the used weight, combined green interval estimates have been calculated separately for two different weights $w_{red} = \{1/8, 1\}$ on the through movement $m = 2$ at intersection Ingolstädter Straße / Neuherbergstraße. Referring to the simulated traffic actuated intersection, shown in Subsection 4.5.2, the author has the assumption that a large number of collected stop seconds in combination with an unadapted weight of w_{red} led again to an overcompensation of the interpolation error which causes a noticeable shift of the green start to a later cycle second.

Unadapted means thereby, that the weighting factor w_{red} is statically and not automatically adapted to movements degree of saturation. Figure 5.1 illustrates the combined green interval estimate, valid for multiple hours that have been grouped together by the result clustering and time assignment algorithm. The vertical dashed lines indicates start and end of the green interval. The dashed black line indicates the remaining consistent number of crossing events per cycle second that is free of any contradicting stop seconds. Thus, the red dashed and the green lines represent frequencies of measured stop seconds and crossing events per cycle second. The mentioned assumption



(a) Combined estimation with $w_{red} = 1$



(b) Combined estimation with $w_{red} = 1/8$

Figure 5.1: Combined estimation results at Ingolstädter Straße / Neuherbergstraße

is confirmed because green starts between both variants differ by 5 secs and is shifted in Figure 5.1a to a later cycle second.

The resulting signal timing parameters for all movements and hours are very extensive and thus have been attached to Appendix D.1 and D.2.

Table 5.8: Estimation quality per results class for intersection at Ingolstädter Straße / Frankfurter Ring, id=13

					QKZ class			
$class$	$mn(QKZ_1)$	$mn(QKZ_2)$	$sd(QKZ_1)$	$sd(QKZ_2)$	A	B	C	$\sum$
1	0.73	0.20	0.30	0.12	76	16	11	103
3	0.77	0.20	0.29	0.20	217	16	57	290

5.1.4 Evaluating Estimated Signal Timing Parameters

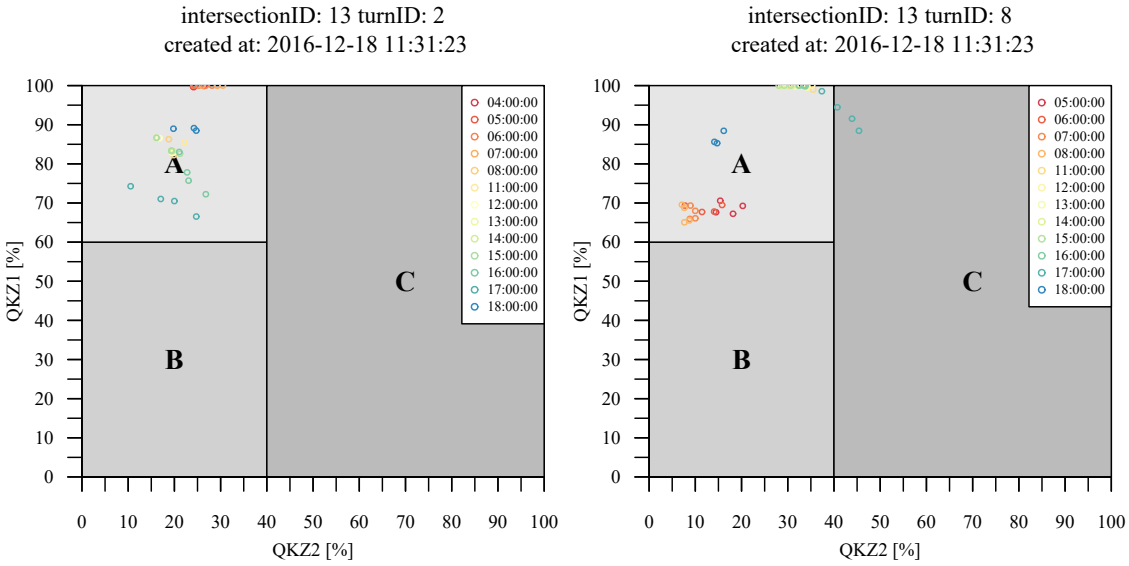
Intersection: Ingolstädter Straße / Frankfurter Ring

Next, results at intersection Ingolstädter Straße / Frankfurter Ring are investigated in an advanced manner. The assessed estimates are differentiated with respect to their result classes $class$, introduced in Section 4.8. Each $class$ allows to distinguish whether an estimate has been proved by at least one conflicting movement. The review of the QKZ values and the result classes in Table 5.8 demonstrate that a higher result class is not a guarantee for a good estimate. Thus, even $class = 3$ contains a relatively high number of low quality estimates. In total approximately 75 % of all estimates are within the desired QKZ class A.

The next stage of the result analysis covers QKZ measures as well as box-plots that present the absolute deviations in seconds between the ground truth and the estimated green intervals over the workday period. Time of day is measured in the UTC time zone (Coordinated Universal Time) and denoted in the 24-hour clock time format. Due to the rich amount of information, author decided to illustrate in this Section exemplary QKZ figures for the through movements $m = \{2, 8\}$ as well as the turning movements $m = \{6, 10\}$. The remaining figures and box-plots are attached in Appendix D.1 of this thesis. Most important and general findings are itemized in the following paragraphs, whereas statements regarding absolute errors for the start and end of green refer always to the measured median deviation between the ground truth and the estimate.

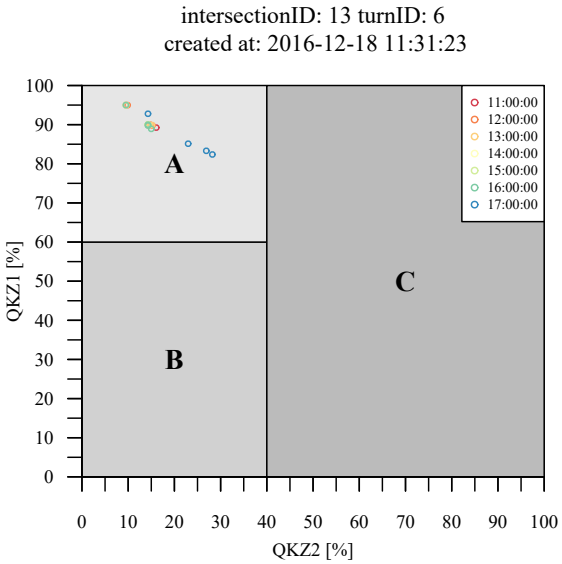
- **Through movements:**

  The movements $m = \{2, 8\}$ show in general a better result availability over the time of day than turning movements which seem to be related to the trajectory volumes, documented in Appendix D.1.2. The movement $m = 2$ shows in total the best estimation quality at this tested intersection. The absolute median deviation for start and end of green is always is less than or equal to 10 secs. In many cases the start and the end of green has been estimated only approximately 5 secs too late.

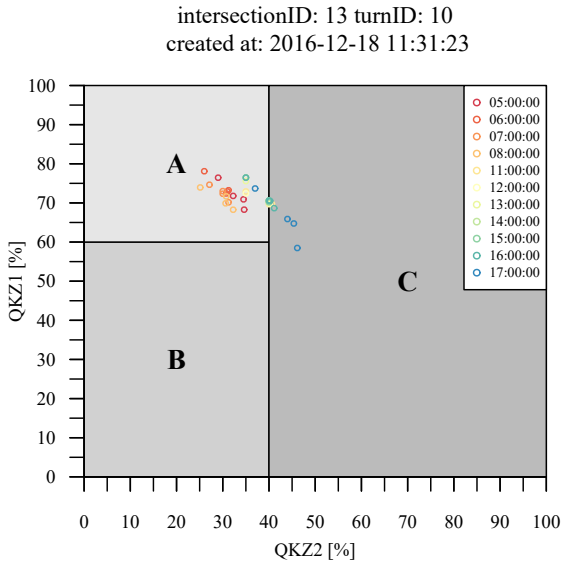


(a) QKZ for movement $m = 2$

(b) QKZ for movement $m = 8$



(c) QKZ for movement $m = 6$



(d) QKZ for movement $m = 10$

Figure 5.2: QKZ measures at intersection Ingolstädter Straße / Frankfurter Ring

The movement $m = 8$ seems to have in comparison to $m = 2$ a much higher degree of traffic actuation, indicated by larger interquartile ranges of the box-plots (see Appendix D.1.3). In general, a different degrees of actuation on single movements within the same phase can be explained by the fact that the green interval duration could get adapted with respect to movements' measured traffic demand. Furthermore, it should be remarked that a high degree of actuation on one particular phase does not need to cause the same degree of actuation on all other phases

at the same intersection, because a following phase could compensate an extended or shortened previous phase in order to achieve a good coordination. The deviation for the start and end of green at movement $m = 8$ is mostly within an absolute median error of 5 secs. In the midday period, the end of green is approximately 10 secs too late. The most plausible reason for this kind of errors is the fact that estimation methodology infers in tendency the maximal available green interval for each movement per hour. As the data analysis covers a very long time period, extreme situations in which the green interval has been extended or shortened, have very likely been observed by trajectories. If these situations occurs frequently enough, the estimated green interval cover these extrema and interpret it as valid interval boundaries.

- **Turning movements:**

  The movements $m = \{4, 6\}$ are the turning movements with most best estimation quality at the tested intersection which is very astonishing because these movements indicate the lowest trajectory volumes, documented in Appendix D.1.2. Thus, the result availability on these two movements is the desired of the developed clustering and time assignment algorithm. The estimation quality of movement $m = 6$ is superior to movement $m = 4$ whereas, the absolute median error for the start and the end of green at movement $m = 6$ is always is less than or equal to 5 secs. The start and end of green at movement $m = 4$ has an absolute median error which is mostly less than or equal to 7 secs. Both movements seems to be influenced rarely by the traffic actuation which can explain the good estimation quality.

  The movements $m = \{10, 12\}$ have been estimated with exactly the same start of green and therefore shows the same error, which is for the absolute median error approximately 5 secs. Despite the fact that both movements are controlled by the very same signal group, the estimated end of green quality at movement $m = 10$ is superior to $m = 12$ and has an absolute median error of less than 8 secs. In comparisons to movements $m = \{4, 6\}$, movements $m = \{10, 12\}$ seem to have a slightly higher degree of actuation, indicated by box-plot outliers that are documented in Appendix D.1.3.

  The movement $m = \{1, 9\}$ have the lowest estimation quality at this tested intersection. The estimated start of green on movement $m = 9$ has mostly an absolute median error of less than 3 secs, whereas the end of green is systematically underestimated by approximately 11 secs, which can be explained as follows. Movement $m = 9$ is a permissive left-turn and is controlled together with movement $m = 8$ in the very same signal group. If this signal group switches to green, only a small proportion of

the left-turning vehicles can cross the stop bar and need thus to wait behind it, till oncoming traffic has fully cleared the intersection or time headways are long enough to allow these vehicles a safe turn. Thus, a crossing event driven green interval estimation on permissive left-turn movements is only able to deduce the complete green interval, if the oncoming traffic is sparse enough and shows therefore wide time headways.

The movement $m = 1$ has also a very interesting finding. The estimated start of green is mostly delayed by approximately 3 secs, whereas the end of green seems be delayed by more than 30 secs in the afternoon. This movement is signalized in two different phases, once in permissive and once in protected mode which means that this movement has two green intervals during the complete cycle time. Due to traffic actuation of this intersection, the simple green interval estimation methodology detects in the afternoon one contiguous long green interval, depicted in Figure 5.3. As the only conflicting movement $m = 5$ could not be evaluated for reasons of the map matching (partially grade separated intersection), no verification and adjustment procedure with conflicting movements can be applied and the output of the simple green interval estimation methodology represents already the final result. If an estimate overlaps with two potential ground truth intervals, it is always compared against the ground truth interval that maximizes the overlapping time. In this particular case, the green interval in the permissive phase shares most

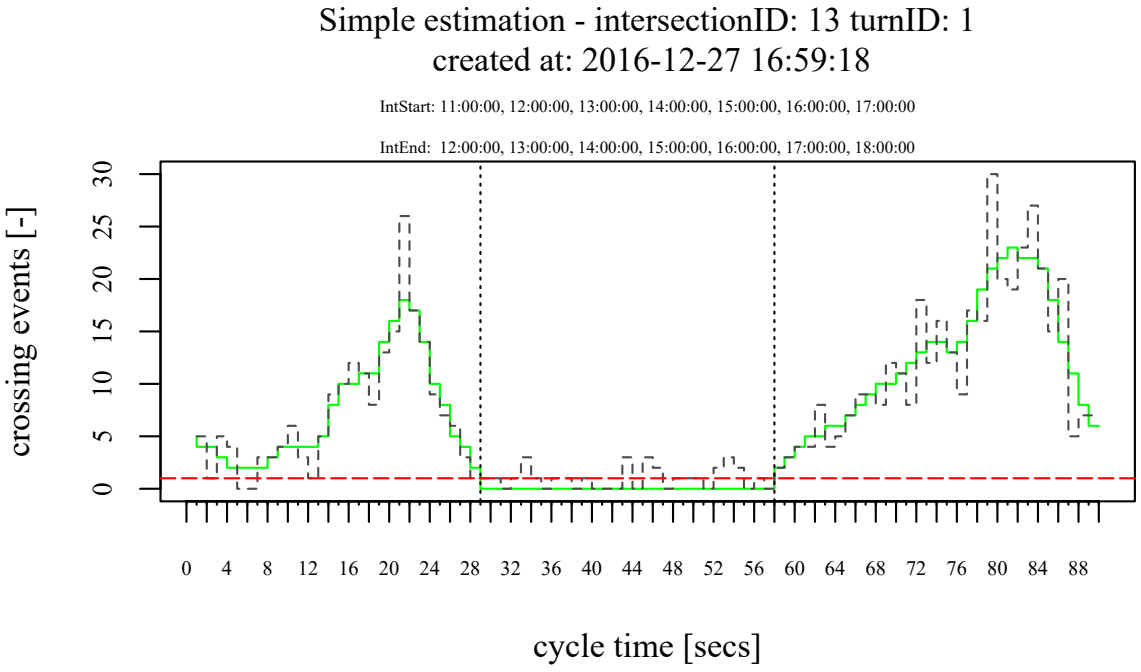


Figure 5.3: Simple estimation of movement $m = 1$ at intersection Ingolstädter Straße / Frankfurter Ring

seconds of the cycle time with the estimate. Thus, the green start of the permissive phase fits very well to the estimated start but not to the estimated end of green, which explains the high deviation.

- **General aspects:** The start of green is in tendency better estimated than the green end. In most cases, the green end is overestimated, which is caused by the effect that the signal timing estimation methodology detected the maximal available green interval for each movement. The one year trajectory dataset covers thereby very likely extreme situations of traffic actuation that occurred only rarely in the five days ground truth comparison period. Nevertheless, even at this very challenging traffic actuated intersection, the established approach was able to estimate reliably the cycle length as well as the start and end of green mostly within an absolute median error of 5-10 secs. The deviations between the ground truth and the estimated green interval at permissive left-turn movements, such as $m = 9$, has been identified as a method respectively data inherent effect.

  The author assessed in addition the green interval estimates with and without the developed result clustering and time assignment algorithm. Without this approach, estimates would have been only available on movement $m = \{2, 8\}$. The clustering caused thereby no noticeable changes in the green interval estimation quality of movement $m = \{2, 8\}$. In fact, the clustering approach has been the fundamental mechanism to allow the inference of green interval estimates on movements with less trajectory volumes. Appendix D.2 gives an overview of the found cluster structure and all corresponding green interval estimates used in the presented ground truth comparison.

Intersection: Ingolstädter Straße / Neuherbergstraße

The second intersection of interest in the calibration dataset is again traffic actuated, whereas the degree of actuation seems to be very small and mainly made the purpose of bus prioritization. The tested intersection allows thereby only the inference of signal timing parameters for movement $m = \{1, 2, 8, 12\}$. All assessed estimates have again been distinguished with respect to their achieved result classes $class$, summarized in Table 5.9. The overall estimation quality is superior to the first intersection and the third result class covers only estimates of the best QKZ class. As the total number of estimated movements it less than for the first intersection, this section covers all available QKZ figures for the through movements $m = \{2, 8\}$ as well as the turning movements $m = \{1, 12\}$. All remaining box-plots that assess the absolute deviations for the green start and end are again attached in Appendix D.2. Most important and general findings are itemized in the following paragraphs.

Table 5.9: Estimation quality per results class for intersection at Ingolstädter Straße / Neuherbergstraße, id=2

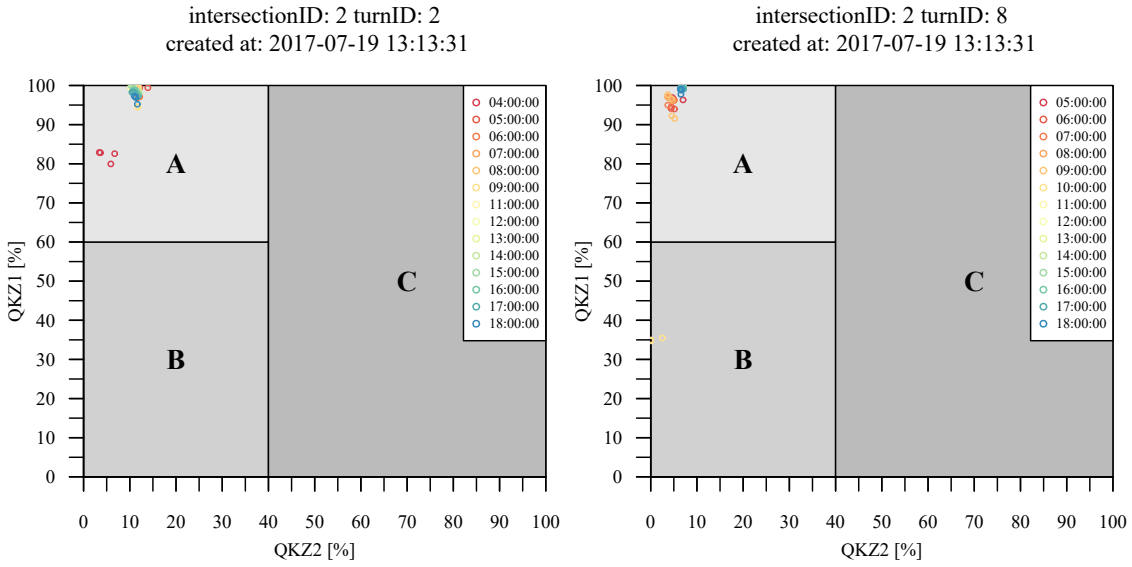
					QKZ class			
$class$	$mn(QKZ_1)$	$mn(QKZ_2)$	$sd(QKZ_1)$	$sd(QKZ_2)$	A	B	C	$\sum$
1	0.84	0.14	0.09	0.03	56	4	0	60
3	0.96	0.04	0.07	0.04	156	0	0	156

- **Through movements:**

  The trough movements $m = \{2, 8\}$ show a very high estimation quality for each available hour. The absolute median error for the start of green on both movements is approximately 3 secs, whereas the end of green shows similarly to the first intersection the tendency to be overestimated by approximately 5-6 secs. The estimates of movement $m = 8$ show a drop of QKZ_1 for 10:00 to 11:00 h (UTC). This effect has been investigated and is reasoned by a too short estimated green interval, caused by too less trajectory volumes.

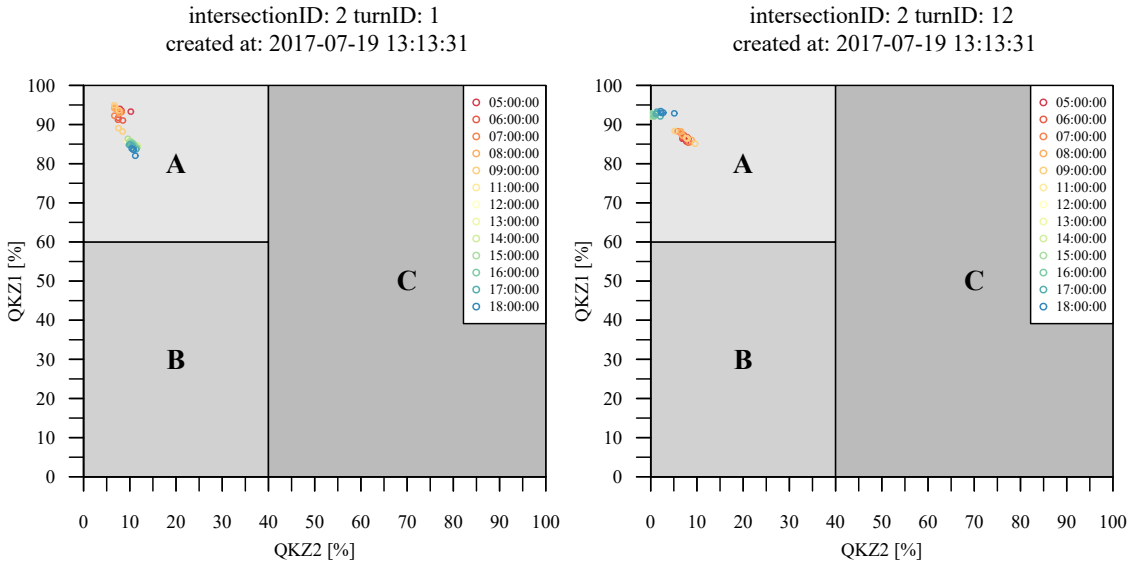
- **Turning movements:**

  The turning movements $m = \{1, 12\}$ show in comparison to movements $m = \{2, 8\}$ slightly lower QKZ_1 measures. The green start on both movements shows an absolute median error of approximately 3-5 secs, whereas the end of green shows especially on movement $m = 1$ an absolute median error up to 6 secs. The relatively good estimation quality for the permissive left turning movement $m = 12$ can be explained by the negligible traffic demand on the movements $m = \{4, 5\}$ that allows the left turning vehicles to pass the intersection during the complete green interval. A closer look at the result availability over the time of day shows that some hours, such as 09:00 and 10:00 h (UTC), are missing. This effect is again reasoned by temporarily too low trajectory volumes.



(a) QKZ for movement $m = 2$

(b) QKZ for movement $m = 8$



(c) QKZ for movement $m = 1$

(d) QKZ for movement $m = 12$

Figure 5.4: QKZ measures at intersection Ingolstädter Straße / Neuherbergstraße

- **General aspects:**

  The second tested intersection confirms some of the already found tendencies. Again, the green start could be better estimated than the end of green which suffers typically from an overestimation of approximately 5-6 secs. The nearly fixed-time controlled intersection allows thus the developed methodology to infer the green intervals mainly within the desired QKZ class A. Without the application of the developed result clustering and time assignment algorithm, estimates would mainly have been available for movements $m = \{2, 8\}$.

5.1.5 Conclusion

The developed signal timing estimation methodology has successfully been proved at two real-world intersections. Both tested intersections show thereby very different and challenging degrees of traffic actuation. Keeping in mind that all estimates could be deduced from less than 30 k trajectories per intersection and approximately 83 % of all calculated estimates are within the QKZ class A (see Table 5.10), the author concludes that the calibration procedure was successful. Nevertheless, it is not negligible that approximately 11 % of all estimated green intervals are within QKZ class C which is according to author's opinion not acceptable but also not very easy to resolve.

The detailed result analysis and especially the absolute deviations for the start and end of green (see Appendix D) confirms that in most cases a delayed green end has caused the high QKZ_2 values. The problem is easy to explain, because the green interval estimation is mainly based on a histogram analysis of stop bar crossing events and the fundamental principle assumes a fixed-time controlled intersection. The methodology tries to infer a minimum frequency level that needs to be exceeded, whereas cycle seconds need to form a contiguous time interval that is used as the estimated green interval. The method assumes that the found time interval is a reasonable estimate of the true fixed-time green interval. If the degree of actuation is increased, the estimated green interval could cover crossing events that occurred relatively infrequent and only in very extreme traffic demand situations. The one year time period which has been exploited renders these extreme events often enough and the used histogram analysis interprets these extreme events as valid green interval boundaries. A traffic signal state prediction that makes

Table 5.10: QKZ class distribution of the calibration dataset

	QKZ Class		
	A	B	C
absolute proportions	505	36	68
relative proportions	82.9 %	5.9 %	11.2 %

use of these extreme values can cause a too early estimated start and a delayed estimated end of green.

Thus, one has the objective to deduce the degree of actuation and to select for traffic actuated movements a more representative green interval. As the green interval estimation procedure considers only the exploitation of an one-dimensional information in a histogram analysis, it is hardly possible to infer reliably the degree of actuation for an isolated movement. To get at least a rough finding of the green interval sensitivity, an quantile-based outlier analysis of crossing events within the found green interval could be established. Figure 5.5 shows this idea in which the 5th and 95th percentiles of crossing events within the originally found green interval are labeled as more representative interval boundaries. The vertical blue lines indicate the percentile-based boundaries, whereas the vertical black dashed lines represent the original interval boundaries as described in Subsection 4.5.1.

On the one hand, the crossing event outliers could naturally be caused by the traffic demand distribution itself and the originally found green interval (dashed black lines) is a good estimate of the true green interval, which belongs to a fixed-time controlled intersection. On the other hand, the very same situation could also be explained by a highly actuated movement at which crossing events close to the interval boundaries occurred less likely because interval lengths and positions within the cycle time have been actuated dynamically. These two situations can not be differentiated from the perspective of an isolated histogram analysis. However, a quantile-based outlier analysis

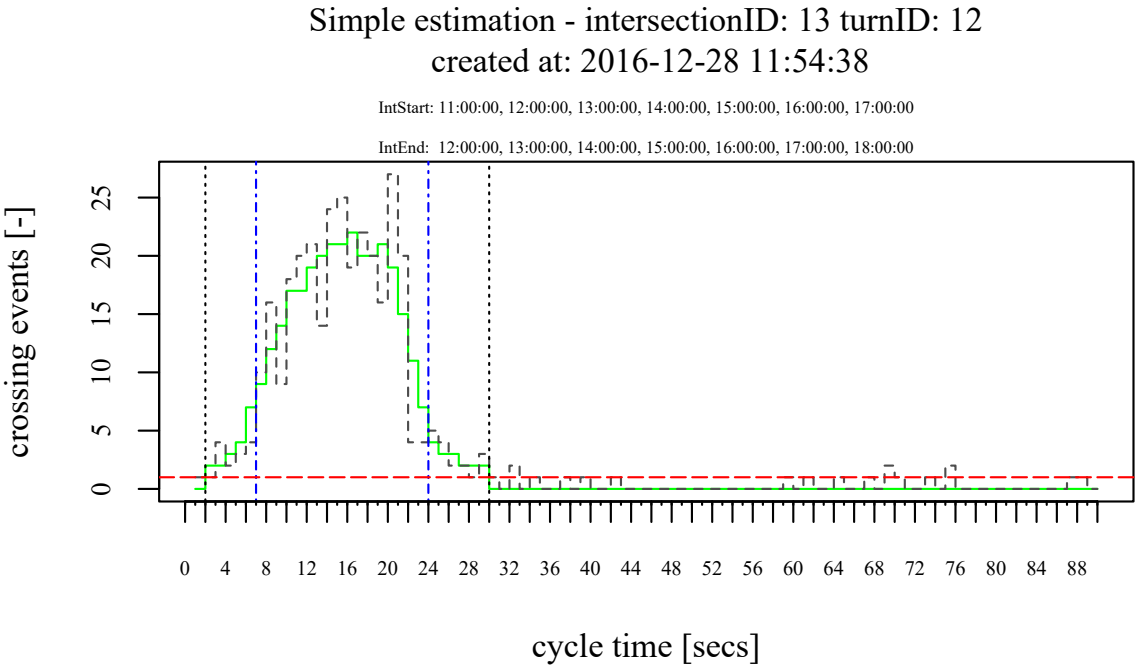


Figure 5.5: Simple estimation of movement $m = 12$ at intersection Ingolstädter Straße / Frankfurter Ring

could be beneficial to assess the interval's sensitivity. A more reliable evidence to distinguish between an actuated or fixed-time controlled intersections can be obtained by comparing the green intervals of conflicting movements.

The thesis presents with the introduced green interval verification and adjustment procedure (see Section 4.7) already an approach that tries to find in cases of traffic actuation a more representative green interval by verifying and adjusting green intervals of conflicting movements. In case of a fixed or nearly fixed-time controlled intersection, the green intervals of conflicting movements are already well separated and match their necessary intergreen times. The verification and adjustment procedure is of minor priory and rarely used to solve conflicting green intervals. An increasing degree of actuation causes longer periods of overlapping conflicting green intervals that needs to be solved by the verification procedure. Thus, an evaluation of the estimated green intervals before and after this verification procedure could be used to calculate the mean difference for start and end of green over all estimates. If for instance this mean difference exceed a particular threshold, it could be assumed that the intersection has reached a degree of actuation and green interval estimates that are not verified by conflicting movements should be rejected. Such a threshold needs even to be calibrated which could not be managed in this thesis. However, the developed estimation methodology could be improved by enhancing the verification procedure in order to achieve less but therefore maybe more representative estimates.

The enhanced verification procedure calculates for each estimated green interval of a given hour at movement i the ratio p_i that describes the quotient between checked conflicting movements n_i to all potential conflicts N_i that could occur. If p_i is greater than or equal to p_{min} estimates are assumed to be reliable enough for further usage. Because the number of movements with plenty trajectory volumes is quite small, the author decided to use $p_{min} = 0.5$. That means each estimated green interval needs at least to be verified by 50 % of its conflicting movements. Nevertheless, a drawback of this concept is the unnecessary rejection of estimates in cases, in which the intersection is in reality fixed-time controlled and/or only a few movements allowed the inference of green intervals.

According to this excursus, the main requirement of a representative green interval estimate, especially on an actuated intersection, is a better data provisioning that allows to verify and adjust each found green interval against all its conflicting movements. This would also explain the effect that the nearly fixed-time controlled intersection at Ingolstädter Straße / Neuherbergstraße was able to produce estimates with a high estimation quality although a few movements showed up available estimates. In contrast thereto, the highly traffic actuated intersection at Ingolstädter Straße / Frankfurter Ring showed up estimates only very temporary on 8 out of 12 movements which causes partially a low estimation quality.

5.2 Validation

5.2.1 Cycle Length Estimation

After this first successful real-world test, the validation has the objective to demonstrate that the calibrated and enhanced methodology is capable to infer signal parameters at different intersections within a very similar or even a higher quality. Therefore, two new intersections likewise in Munich (Germany) are selected. Most important properties and signal phasing information are again summarized by Table 5.11 and Table 5.12.

Especially the intersection at Dachauer Straße / Georg-Brauchle-Ring has been selected to assess the methodology under very challenging conditions. The intersection of interest includes thereby a prioritized and frequently used tramway at which every 10 min trams pass by. It can be assumed that the regular occurrence of trams could cause a very irregular trajectory pattern which could interfere the green interval estimation procedure and maybe also the cycle length estimation. Referring to the calibration procedure, the validation starts again by applying the cycle length estimation from Section 4.4. A historical trajectory dataset of the very same time period (03/2014 to 03/2015) is again used to figure out the typical cycle length for the workday period with an hourly time resolution by testing potential cycle length in a range from 30 to 120 secs. The cycle length estimation has been repeated twice by testing the most common backward calculation methods in Germany (RRV and RRV 1.1., see Appendix F).

Only the backward calculation method "RRV" allowed again the inference of the cycle length $C = 70$ secs at the intersection Frankfurter Ring / Ungererstraße. Again, trajectory volumes in the early morning and late evening are not dense enough to estimate the cycle lengths with the required significance level of $\alpha = 0.005$. The presence of the tram does not seem to interfere the cycle length estimation procedure at intersection Dachauer Straße / Georg-Brauchle-Ring.

Table 5.11: Fact sheet: Dachauer Straße / Georg-Brauchle-Ring

Intersection id:	1
Major road:	Georg-Brauchle-Ring
Signal control type:	actuated
Signal phase structure:[1]	Ph1=5,11,12 Ph2=3,9 Ph3=2,8
Public transport:	Bus and tramway
Daily traffic volume:[2]	40/30
Intersection layout:	4 arms
Ground truth period:	07-11/07/2014
Specific features:	channelized right-turns, movement 6 is prohibited

[1] movement numeration schema (see Section 3.2)
[2] units in veh/day for major/minor road

Table 5.12: Fact sheet: Frankfurter Ring / Ungererstraße

Intersection id:	16
Major road:	Frankfurter Ring
Signal control type:	actuated
Signal phase structure:[1]	Ph1=1,2,3,7,8,9 Ph2=11,12 Ph3=4,5,6
Public transport:	Bus
Daily traffic volume:[2]	37/32
Intersection layout:	4 arms
Ground truth period:	07-18/07/2014
Specific features:	movement 10 as channelized right-turn

[1] movement numeration schema (see Section 3.2)
[2] units in veh/day for major/minor road

Table 5.13: Estimated (C) and ground truth cycle length (C_{true}) at Dachauer Straße / Georg-Brauchle-Ring

h UTC	m	C [sec]	C_{true} [sec]	A^2 [-]	p_{value} [-]	total vol. [veh/h]
00:00:00	8	-	70	-	-	98
01:00:00	5	-	70	-	-	84
02:00:00	2	-	70	-	-	68
03:00:00	2	-	90	-	-	106
04:00:00	2	90	90	32.57	5.22e-06	185
05:00:00	2	90	90	24.96	3.14e-06	329
06:00:00	2	90	90	27.28	2.53e-06	400
07:00:00	2	90	90	28.01	2.33e-06	463
08:00:00	2	90	90	47.48	1.99e-06	522
09:00:00	2	90	90	54.38	2.16e-06	474
10:00:00	2	90	90	88.38	2.01e-06	527
11:00:00	2	90	90	132.08	1.95e-06	537
12:00:00	2	90	90	91.45	2.24e-06	543
13:00:00	8	90	90	94.43	2.63e-06	543
14:00:00	8	90	90	95.72	2.64e-06	588
15:00:00	8	90	90	106.79	2.21e-06	651
16:00:00	8	90	90	129.13	1.89e-06	804
17:00:00	2	90	90	173.57	1.38e-06	827
18:00:00	2	90	90	157.78	1.50e-06	683
19:00:00	8	90	90	61.26	3.53e-06	450
20:00:00	8	90	90	43.58	5.13e-06	366
21:00:00	8	90	90	44.43	5.17e-06	342
22:00:00	8	-	90	-	-	238
23:00:00	8	-	90	-	-	151

Table 5.14: Estimated (C) and ground truth cycle length (C_{true}) at Frankfurter Ring / Ungererstraße

h UTC	m	C [sec]	C_{true} [sec]	A^2 [-]	p_{value} [-]	total vol. [veh/h]
00:00:00	12	-	70	-	-	59
01:00:00	12	-	70	-	-	57
02:00:00	12	-	70	-	-	58
03:00:00	8	-	70	-	-	406
04:00:00	8	-	70	-	-	736
05:00:00	8	90	90	87.76	3.97e-06	874
06:00:00	8	90	90	66.18	3.47e-06	913
07:00:00	8	90	90	86.36	3.05e-06	929
08:00:00	8	90	90	110.41	2.67e-06	1043
09:00:00	8	90	90	139.33	2.40e-06	968
10:00:00	8	90	90	230.40	1.60e-06	1016
11:00:00	8	90	90	343.21	1.20e-06	1054
12:00:00	8	90	90	380.25	1.15e-06	1062
13:00:00	8	90	90	412.71	1.10e-06	1053
14:00:00	8	90	90	398.69	1.21e-06	1030
15:00:00	8	90	90	433.23	1.08e-06	1141
16:00:00	8	90	90	382.17	1.14e-06	1122
17:00:00	8	90	90	383.10	1.01e-06	1229
18:00:00	8	90	90	168.31	1.73e-06	804
19:00:00	12	70	70	25.97	2.91e-06	337
20:00:00	12	70	70	29.81	5.00e-06	227
21:00:00	12	70	70	59.33	3.64e-06	229
22:00:00	12	70	70	28.75	6.00e-06	151
23:00:00	12	-	70	-	-	55

5.2.2 Evaluating Estimated Signal Timing Parameters

Intersection: Dachauer Straße / Georg-Brauchle-Ring

Next, the green intervals have been estimated with an hourly time resolution for the workday period and assessed against the ground truth at the intersection Dachauer Straße / Georg-Brauchle-Ring for two different scenarios. The first scenario incorporates the original green interval verification and adjustment procedure which permits all calculated estimates. The second scenario takes the enhanced concept of Subsection 5.1.5 into account, in which only green intervals that have been verified by at least 50 % ($p_{min} = 0.5$) of their conflicting movements are accepted. All estimated green intervals, the time assignment table and available trajectory volumes are documented in Appendix D.3 of this thesis.

Table 5.15 shows first the assessed results of the original green interval verification procedure. It is visible that the intersection contains approximately 50 % of low quality estimates that belong to the QKZ class C. The previous experience has shown that this low estimation quality is mainly caused by an overestimated end of green at traffic actuated movements. According to this consideration, the enhanced verification procedure should eliminate these estimates, which is confirmed by Table 5.16.

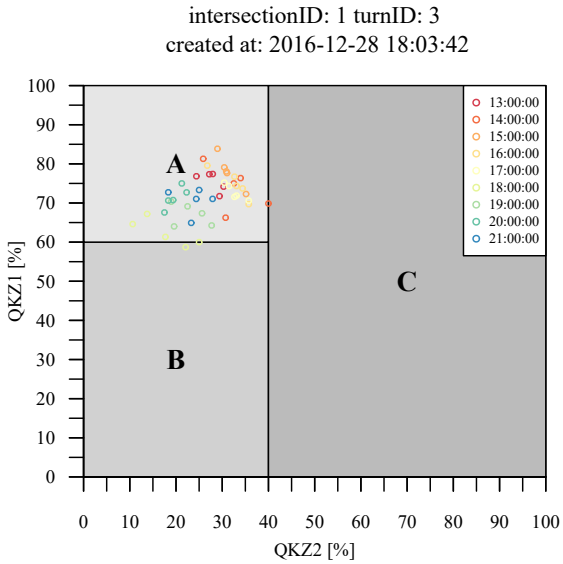
The enhanced verification procedure caused a remarkable reduction of the mean QKZ_2 from 32 to 25 %. Furthermore, only a small number of 20 green interval estimates are remaining within the QKZ class C, which should be investigated in an advanced manner in the following paragraphs. As a negative side effect of the enhanced verification procedure, approximately 60 % of all green interval estimates have been filtered, which is plausible and mainly caused by a lower result availability till 13:00 h (UTC). Up to this time of the day, mainly movements $m = \{2, 5, 8, 11\}$ show estimates, whereas results on the important left-turning movements are missing. All in all, only four movements $m = \{3, 5, 9, 11\}$ are satisfying temporarily the constraint of the enhanced

Table 5.15: Estimation quality per results class at intersection Dachauer Straße / Georg-Brauchle-Ring, id=1, $p_{min} = 0$

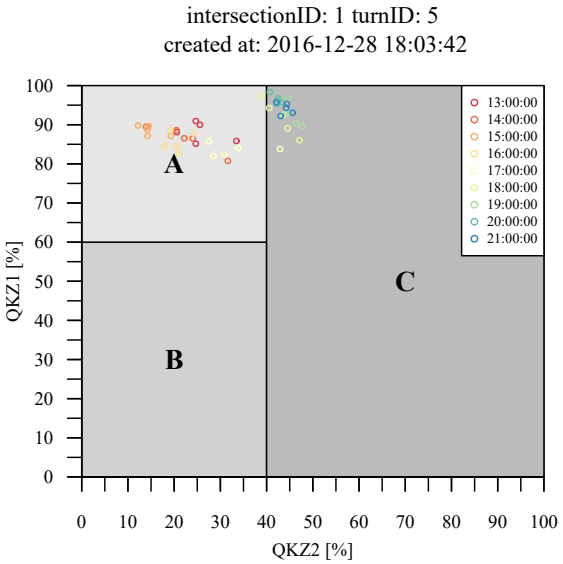
					QKZ class			
class	$mn(QKZ_1)$	$mn(QKZ_2)$	$sd(QKZ_1)$	$sd(QKZ_2)$	A	B	C	$\sum$
1	0.70	0.04	0.32	0.14	4	0	1	5
3	0.86	0.08	0.32	0.13	263	2	130	395

Table 5.16: Estimation quality per results class at intersection Dachauer Straße / Georg-Brauchle-Ring, id=1, $p_{min} = 0.5$

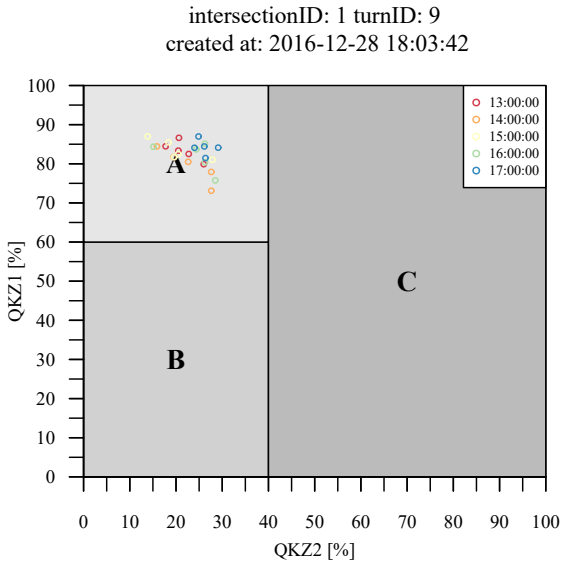
					QKZ class			
class	$mn(QKZ_1)$	$mn(QKZ_2)$	$s(QKZ_1)$	$s(QKZ_2)$	A	B	C	$\sum$
1	0.83	0.09	0.25	0.10	138	2	20	160



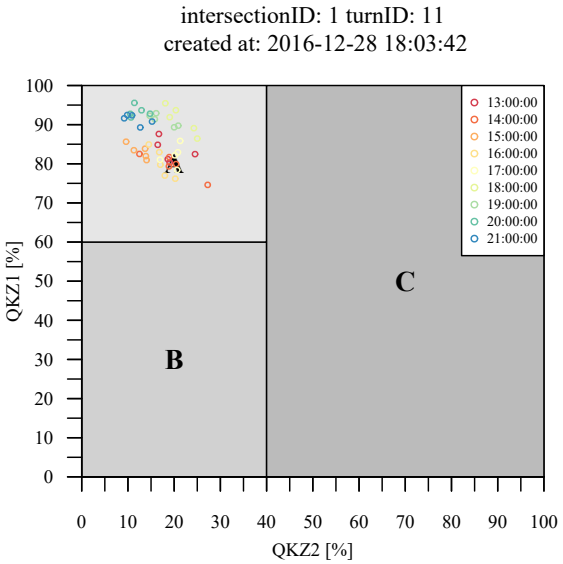
(a) QKZ for movement $m = 3$



(b) QKZ for movement $m = 5$



(c) QKZ for movement $m = 9$



(d) QKZ for movement $m = 11$

Figure 5.6: QKZ measures at intersection Dachauer Straße / Georg-Brauchle-Ring

verification procedure, whereas the reaming results achieve a remarkable estimation quality, illustrated by QKZ measures in Figure 5.6. A detailed list that contains all available estimates for a given hour and movement has been attached in Appendix D.7. The most important and general findings are again itemized in the following paragraphs.

- **Through movements:**

  Figure 5.6 and the previous explanations state that the through movements $m = \{5, 11\}$ could be temporarily verified by 50 % of their conflicting movements, whereas the estimation quality of movement $m = 11$ is superior to $m = 5$. The box-plots in Appendix D.3.3 prove that the absolute median deviation of the start and the end of green at movement $m = 11$ is less than or equal to 5 secs. Some outliers and temporarily wide interquartile ranges in box-plots indicate that the intersection of interest seems to have in comparison to all other tested intersections a higher degree of actuation, which can be explained by the existing public transport prioritization. It is conspicuous that the estimated end of green at movement $m = 5$ suffers after 18:00 h (UTC) from a strong delay that ranges up to approximately 13 secs. This phenomenon has been analyzed and it can be stated that missing results on conflicting movements are not the main reason for this problem.

  The author assumes that the found phenomenon is related to the result clustering and time assignment algorithm which tries to find hours of similar green intervals on all movements. The found hours are aggregated to achieve on movements with less trajectory volumes a sufficient data density. The time assignment table in Appendix D.3.1 shows that after 18:00 h (UTC) one large cluster with the abbreviation "2_2_90_6" is found by the help of movement $m = 2$. The trajectory volumes, also documented in Appendix D.3.2 renders between 17:00 and 18:00 h (UTC) a significant drop in the traffic demand, whereas the volumes on movement $m = 5$ are decreasing much faster than on movement $m = 2$. While the traffic demand and therefore green intervals at movement $m = 2$ remain after 18:00 h (UTC) relatively constant, the green intervals at movement $m = 5$ are adapted to the changed traffic demand situation, which can be proved by a statistical analysis of ground truth traffic signal states (see Appendix D.3.4). The analysis confirms that the green start of movement $m = 5$ is shifted between 17:00 h and 22:00 h (UTC) from approximately second 31 to second 21 of the cycle, whereas on movement $m = 2$ no substantial changes in green interval position or interval length can be determined. Thus, the error cause of the delayed end of green at movement $m = 5$ is an insufficient adaption of the estimated green interval to the rapidly changing traffic demand over the time of day.

- **Turning movements:**

  The estimates on the left-turning movements $m = \{3, 9\}$ show similar to the through movements a remarkable estimation quality and all results are within the desired QKZ class A. The absolute median deviations for the start and the end of green are for all available estimates less than or equal to 5 secs.

Intersection: Frankfurter Ring / Ungererstraße

The real-world validation of this chapter finishes with a final result analysis at the intersection Frankfurter Ring / Ungererstraße. This last validation has been carried out for the original and the enhanced green interval verification and adjustment procedure. For a detailed view on all estimated green intervals, the time assignment table and all available trajectory volumes, the thesis refers to Appendix D.4.

The comparison of assessed estimation results from Table 5.17 and Table 5.18 demonstrates that the intersection of interests does not seem to benefit from the more strict verification procedure, which could be interpreted as first evidence that most estimates are already verified by their conflicting movements. Thus, the relative large trajectory volumes and the good result availability allowed in total the estimation of green intervals for the movements $m = \{1, 2, 5, 6, 8, 11, 12\}$. Due to the large number of available results, this paragraph presents only QKZ plots for the through movements $m = \{2, 5, 8, 11\}$. The most important findings on all movements are conclusively itemized in the following paragraphs. A detailed overview of all estimates and the box-plot for the estimation quality are attached in Appendix D.4.3.

- **Through movements:**

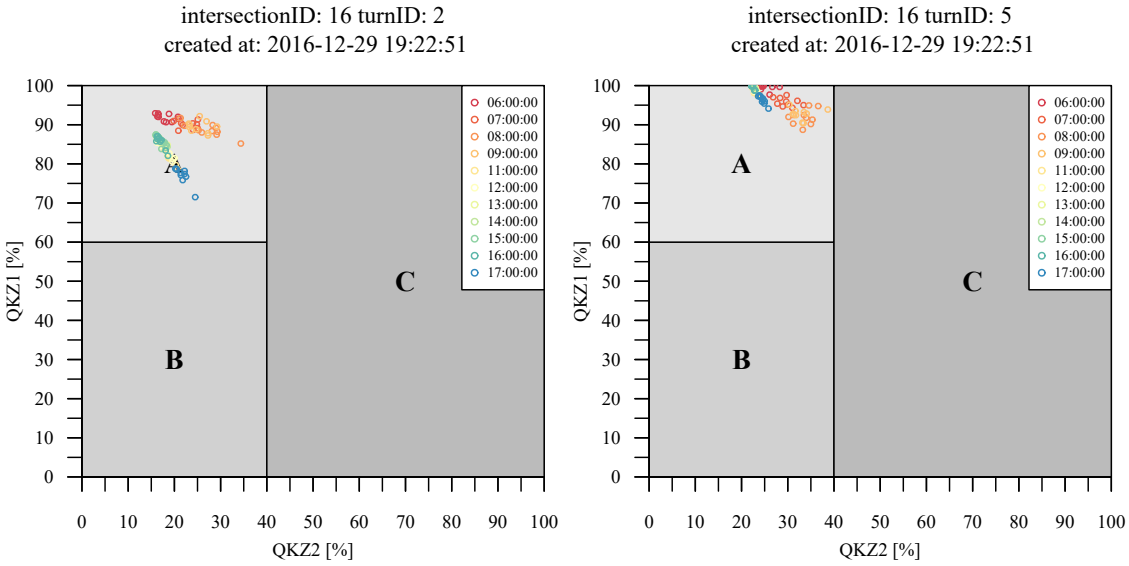
  All trough movements $m = \{2, 5, 8, 11\}$ show in general a relatively high estimation quality and nearly all available estimates are within QKZ class A. The movements $m = \{5, 11\}$ are clearly superior to the movements $m = \{2, 8\}$. A more detailed box-plot inspection shows that the green intervals of movements $m = \{2, 8\}$ suffer from a systematic offset of approximately -5 secs (see Appendix D.4.3). Moreover, the same results confirm that the negative delay is getting smaller between 11:00 to 16:00 h (UTC) for movement $m = 8$, whereas the estimated green interval stays in fact constantly. This can be interpreted as an evidence that the clustered estimates are again not always well adapted to the true

Table 5.17: Estimation quality per results class at intersection Frankfurter Ring / Ungererstraße, id=16, $p_{min} = 0$

					QKZ class			
$class$	$mn(QKZ_1)$	$mn(QKZ_2)$	$sd(QKZ_1)$	$sd(QKZ_2)$	A	B	C	$\sum$
1	0.85	0.07	0.20	0.15	35	0	5	40
3	0.90	0.10	0.22	0.09	683	0	47	730

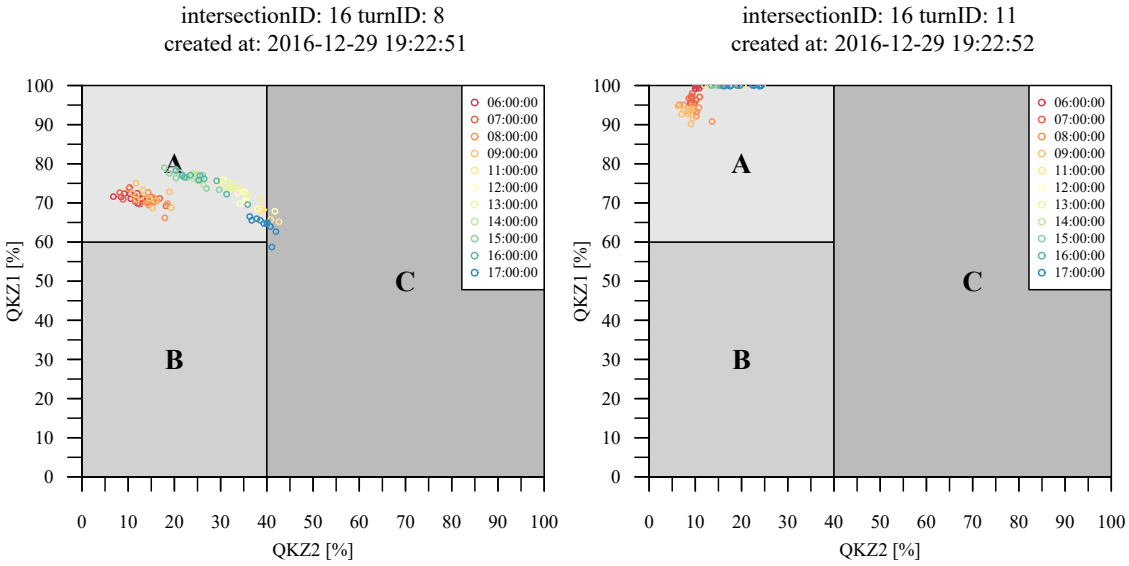
Table 5.18: Estimation quality per results class at intersection Frankfurter Ring / Ungererstraße, id=16, $p_{min} = 0.5$

					QKZ class			
$class$	$mn(QKZ_1)$	$mn(QKZ_2)$	$sd(QKZ_1)$	$sd(QKZ_2)$	A	B	C	$\sum$
1	0.88	0.10	0.24	0.09	573	0	47	620



(a) QKZ for movement $m = 2$

(b) QKZ for movement $m = 5$



(c) QKZ for movement $m = 8$

(d) QKZ for movement $m = 11$

Figure 5.7: QKZ measures at intersection Frankfurter Ring / Ungererstraße

green interval. Thus, slightly changes of the true green interval within 5 secs can not be detected by the developed estimation methodology as it tolerates this deviations during the result clustering procedure. Nevertheless, the estimated start and end of green on all through movements show mainly an absolute median error less than or equal to 5 secs that is temporarily increased on movement $m = \{2, 8\}$ up to 10 secs.

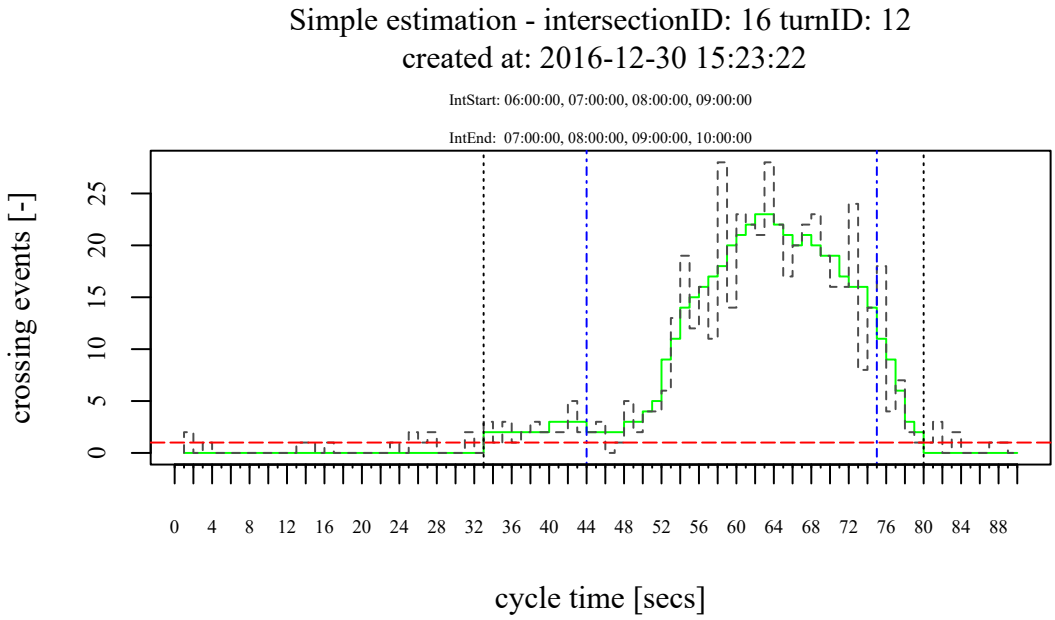


Figure 5.8: Crossing event outliers analysis on movement $m = 12$ between 06:00 - 09:00 h (UTC)

- **Turning movements:**

  The very same or even smaller errors can be deduced on the turning movements $m = \{1, 6, 12\}$. Only the movement $m = 12$ shows temporally in the morning a too early estimated start of green. This phenomenon has been analyzed intensively and can be explained as follows.

  The traffic demand on the left-turning movement $m = 12$ is very intensive and the actuated signal control in charge allows two different green starts during the cycle time, documented by ground truth signal states for this movement in Appendix D.4.4. The histogram-based estimation interpret these two different green intervals as one large green interval, that is illustrated by the black dashed vertical lines in Figure 5.8.

  Even from the view of the verification procedure, the green interval of movement $m = 12$ fits into the cycle time without causing major conflicts with the through movements $m = \{2, 8\}$. Furthermore, the missing estimates on the conflicting movements $m = \{3, 9\}$ would only cause minor improvements of this situation because movements $m = \{2, 3\}$ as well as $m = \{8, 9\}$ are pairwise controlled by the same signal groups and thus an additional verification with green intervals from movements $m = \{3, 9\}$ would cause nearly no beneficial adjustments in the green interval of movement $m = 12$. In such a situation the developed histogram-based methodology has reached its limits.

Table 5.19: QKZ class distribution of the validation dataset

	QKZ Class		
	A	B	C
absolute proportions	711	2	67
relative proportions	91.2 %	0.3 %	8.6 %

To overcome or at least to minimize this effect caused by traffic actuation, the author recommends to establish in general an outlier-based sensitivity analysis of crossing events within the found green interval to prove the representativity of the interval boundaries. This concept is depicted in Figure 5.8. The dashed vertical blue lines therefore indicates the 5th and 95th percentile of crossing events within the originally estimated green interval. Only a relatively small proportion of crossing events (5th percentile) is used to explain a relatively large share (20 %) of the found green interval. By relying on the percentile-based boundaries the start of green would have been shifted into the right direction, from cycle second 34 to second 44 and even the green end would have been pushed closer to a more representative green end from second 77 to 75.

5.2.3 Conclusion

Despite the extreme situation on the left-turning movement $m = 12$, the developed traffic signal state prediction methodology has successfully been proved on two different real-world intersections. Both tested intersections showed thereby again a high degree of traffic actuation. Nevertheless, approximately 91 % of all calculated estimates are within the QKZ class A and thus, the results of the validation dataset outperforms the calibration. The enhanced green interval verification procedure had a remarkable effect on the estimation quality, especially at intersection Dachauer Straße / Georg-Brauchle-Ring. For the time of day, the absolute median deviation of the start and end of green is mostly less than or equal to 5 secs, whereas the start and end of green are in tendency delayed.

Larger deviations between the estimated and the ground truth intervals could have been mainly explained by an insufficient availability of green interval estimates on all conflicting movements. This finding demonstrates that the used FCD source and the derived trajectory volumes are often close to the minimum trajectory volume that is required for the signal timing parameter estimation. Nevertheless, the validation dataset was able to achieve at least the same or even a better estimation quality. It can thus be stated that the validation procedure was successful. Table 5.19 summarizes finally the achieved QKZ class distribution for the validation dataset.

CHAPTER 6

Conclusion

6.1 Summary

In this thesis a novel methodology for the traffic signal state prediction based on vehicle trajectory data has been developed and evaluated. In advance of this development process, the thesis started firstly with a definition of the fundamentals of traffic signal control strategies and introduced briefly the concept of Floating Car Data (FCD). In addition, the latest state-of-the-art of a trajectory-based traffic signal timing parameter and signal state prediction has been carried out to identify needs for further research and development. Referring to this literature review, the thesis identified the need of a profound FCD trajectory analysis in order to understand the data characteristics and trajectory volumes of a real-world, commercially available FCD source. Thus, a trajectory volume analysis has been established at two different major intersections in Munich (Germany). The analyzed one year data aggregate has shown that real-world FCD trajectories are sampled with different time intervals, typically in a range between 15-120 secs. Approximately 50-70 % of the trajectories have been sampled with a sampling interval of 15 secs, whereas the accumulation of all trajectories allowed approximately the daily collection of 15-20 FCD trajectories on the major through movements.

In addition, the literature review discovered the need to develop and describe a comprehensive methodology that covers all major aspects of the trajectory processing in order to infer the signal timing parameters and to estimate traffic signal states based on such sparse and likewise low-frequently sampled, real-world trajectory volumes. Thus, the thesis analyzed, modified and composed a set of already existing algorithms to process real-world trajectories. This set of methods covers the graph modeling, the handling of stopped vehicles, the map-matching and the trajectory filtering. The thesis investigated more detailed the problem of trajectory interpolation and position inaccuracy, which could lead to erroneous stop bar crossing events when working with low-frequently sampled, real-world FCD. In order to reduce the number of

erroneous trajectories beforehand any signal timing estimation, the thesis tested and benchmarked different filter techniques. The benchmark has shown that the application of the Random Forest classification tree algorithm outperformed all competitive methods and allowed thus a very efficient reduction of erroneous trajectories on intersection's through movements.

To allow an efficient and likewise robust traffic signal state prediction, the thesis developed a novel, hypothesis-based cycle length estimation methodology. The most important requirements of this novel approach are recurring signal plans over the same time of day, a mainly constant cycle length within each signal plan and the utilization of a standard backward calculation method. To ensure a sufficient number of available trajectories, the sparse input dataset needs typically to be aggregated for similar hours into a workday and weekend dataset. Thus, the method estimates the typical cycle length for each hour of the aggregated workday dataset by exploiting stop bar crossing events that are projected into the time scale of different cycle length candidates by the help of a modulo operation. Only the application of the correct cycle length causes a concentration of crossing events within the cycle time, whereas the wrong cycle length causes a nearly uniform crossing event distribution. To infer the true cycle length, the fundamental effect of trajectory concentration is exploited by the help of a hypothesis test, which has the objective to find a cycle length candidate that allows to reject an assumed uniform distribution of crossing events with a statistical significance of $\alpha = 0.005$. The concept has been analyzed for a fixed-time controlled intersection by means of a Monte-Carlo-Simulation. Approximately 90-100 crossing events are enough to identify the cycle length, whereas the probability of type II errors remains for this number of crossing events under the level of $\beta = 0.001$. In addition, the established study has shown that the ratio between the green interval and the cycle length for a particular movement plays an important role in the developed cycle length estimation procedure.

The next major research objective of this thesis was the development of a robust green interval estimation methodology that is capable to deduce the green interval for each hour of the aggregated workday dataset by exploiting only some hundred trajectories. In contrast to other approaches, the novel method is primarily data-driven and allows the green interval estimation without assuming a particular crossing event distribution model. Additionally, a rule-based and iterative self-verification process allows to check the consistency of the found green interval which makes the methodology robust when working with small numbers of trajectories. The basic concept of this green interval estimation relies on a crossing event histogram analysis. An iterative classification procedure infers the green interval by analyzing the number of counted crossing events per cycle second. The approach tries thereby to determine automatically a reasonable threshold of required crossing events per cycle second that needs to be exceeded in order to assign a cycle second to the signal state green. A set of predefined data and time related constrains is

used to verify the consistency of the found green interval. Especially in cases in which the complete cycle time is superimposed by uniformly distributed erroneous trajectories, the novel method is capable to adapt automatically the number of required crossing events per cycle second. This first green interval estimation methodology ("simple green interval estimation") has been analyzed in a simulation. The results allowed to state that approximately 200-300 trajectories are necessary to determine the green interval of the tested fixed-time intersection. The start of green has in tendency been estimated 3-5 secs too early, whereas the end of green has been estimated 1-3 secs too late. The mentioned deviations thereby correspond to the 25th and 75th percentiles. The linear trajectory interpolation could be identified as the main reason for the mentioned deviations.

In order to compensate these errors, the thesis introduced another novel concept that tries to make use from stopped vehicles that remain upstream the stop bar. The method evaluates additionally stopped vehicles per cycle second, so called stop seconds, in a histogram analysis. If more crossing events than stop seconds could be sampled per cycle second, the method decides to assign a cycle second to the signal state green. This enhanced concept ("combined green interval estimation") has been even analyzed in a simulation study by considering again a fixed-time intersection. The enhanced concept has in principle the capability to compensate the negative effect of trajectory interpolation, if the method is well calibrated according to movement's degree of saturation.

Another important objective of this thesis was to evaluate the capability to estimate and to predict the traffic signal states at an intersection with a vehicle actuated green interval extension. The study allowed to state that again approximately 200-300 trajectories have been needed to estimate the traffic signal timing parameters for an individual movement. The developed methodology thereby detects and predicts very well the total frame of the green interval, which is typically often used, if traffic demand is nearly stationary and the movement is close to its capacity limit. In the simulated peak hour, the green end could be estimated correctly in 50 % of all predicted cycles. The study underlined the importance of daily and hourly similar traffic demand in order to achieve a good signal state prediction, even at simple actuated intersections.

In addition, the research of this thesis closed the gap of a missing green interval verification procedure that checks potentially overlapping green intervals with respect to their required intergreen times. The developed concept requires ideally for a given hour green interval estimates on all conflicting movements. Since the trajectory volumes on minor roads or turning movements are typically relatively small, the thesis developed moreover a novel cluster-based approach that tries to identify hours with very similar signal timing parameters. This clustering procedure has been tested in a simulation study. On the one hand, the concept allowed to derive estimates on movements

with less trajectory volumes. On the other hand, the clustering can have a negative side effect. Thus, the needed time tolerance for the result clustering could group hours together which are in fact controlled by slightly different signal plans.

The last objective ot this thesis was the calibration and validation of the developed traffic signal state estimation methodology under real-world conditions. Thus, the developed estimation approaches have been tested in a first stage at two major intersections in Munich (Germany) in order to find a parameter setup that optimizes the traffic signal state prediction quality. In a second stage, two different real-world intersections, likewise in Munich, have been analyzed by considering the very same parameter setup. Both, calibration and validation made use from a one year period, commercially available trajectory dataset, obtained from the ADAC (General German Automobile Club). The successful calibration and validation procedure has been accompanied by a set statistical indicators in order to assess and analyze the traffic signal state prediction quality. As expected, nearly fixed-time controlled intersections showed in tendency a better traffic signal state prediction quality than highly actuated ones. Relatively large deviations of more than 5 secs between the estimated and ground truth green intervals could have been mainly explained by missing estimates on important conflicting movements. The profound result analysis has shown that especially actuated intersections ideally require green interval estimates on all conflicting movements in order infer representative results. The study has shown in addition, that a trajectory data driven green interval estimation on permissive left-turning movements is only able to deduce the complete green interval, if the oncoming traffic is sparse enough to allow the left-turning vehicles a safe turn.

An other major finding refers to the result clustering approach. The method has been required to reach even on movements with less trajectory volumes a sufficient sample size to allow the green interval estimation. The approach assumes that the green intervals on all movements within a found time cluster are mostly constant, which may at traffic actuated intersections not always be the case. Thus, the established real-world study identified many different situations in which larger trajectory volumes on all movements over the complete time of day would have probably guaranteed much better green interval estimates. The available trajectory volumes of the used FCD source were especially during the early morning and the late evening insufficient and did not allow to estimate intersection's cycle length with respect to the preassigned level of significance $\alpha = 0.005$.

The results of this real-world study allow to state that for many predicted hours the absolute median deviation for the start and end of green can reach small deviations less than or equal to 5 secs, whereas the start and end of green are in tendency delayed. The most important requirement of a good estimation quality is a sufficient number of at least 200-300 trajectories per hour and over all movements at a signalized intersection. All in all, the real-world study has

shown impressively the high potential of a trajectory-based traffic signal state prediction.

6.2 Outlook on Further Research

The developed estimation methodology of this thesis covers a fairly high variety of interdisciplinary issues that needed to be solved in order to estimate and predict traffic signal states at real-world intersections by exploiting sparse and low-frequently sampled trajectory data. Thus, the methodology of this thesis covered for instance algorithms that refer to the topic of map matching, spatial clustering and classification techniques. The methodology focused especially on the development of a very robust and likewise comprehensible green interval estimation procedure that is capable to operate on small trajectory volumes. During the development, implementation and the later real-world result analysis, different aspects could be figured out that should be investigated in order to improve the prediction quality:

- **Trajectory filtering at turning movements:** The methodology exploits stop bar crossing events as a main information to infer the cycle length as well as the green interval of individual movements. In order to deduced the green interval, the method proposes to filter implausible crossing events that would lead to a too early detected start of green, caused by the linear trajectory interpolation. The results of this thesis proved that the Random Forest Tree classification algorithm (see Section 4.3) is capable to manage this task very efficiently. The algorithm needs to be well trained with simulated crossing events on through movements. Left- and right-turning movements can be signalized in two different modes, i.e. protected and/or permissive. Both modes are naturally affecting the driving dynamics for turning vehicles and thus the classification schema needs know which kind of signalization is used in order to filter trajectories correctly. To improve a classification based data filtering at turning movements, further research is needed that allows to identify reliable the signalization mode of left- and right-turning movements.
- **Classification of movements degree of actuation:** Moreover, the empirical real-world study has shown that a too early estimated green start and a delayed end of green could be caused by an actuated traffic signal control. The study proved that in many situations the degree of actuation seems to be rather a property of a specific movement than the complete intersection. That means, intersection's movements can show different degrees of actuations. So far, the methodology considers green intervals of conflicting movements to prove and adjust estimates in order to find even for actuated movements a reasonable green interval estimate. The developed procedure seems to be very effective but it requires a high

number of estimates on all conflicting movements to ensure a trustworthy estimation result. Thus, it would be beneficial, if a method could be developed which is able to infer for an individual movement its degree of actuation and deduce in return a more representative green interval - without the necessity of estimates on all conflicting movements. As a first reasonable starting point, the thesis proposed in Subsection 5.2.2 to assess each found green interval by a sensitivity analysis, which allows to understand and to interpret the shape of the crossing event distribution especially close to the interval boundaries.

- **Sensing signal plan switchings:** The developed methodology of this thesis determines the signal timing estimates with an hourly time resolution. This process assumes in general that the true signal plan switching takes place close to a full hour, which is at least true for all tested real-world intersections of Chapter 5. However, there may be intersections at which the signal plan switching is executed regularly in between an hour. Such a situation could cause a missing signal timing estimate during the switching hour, if the constrains of data coherence are violated. Moreover, minor changes of the green interval within a few seconds between the previous and the subsequent signal plan are tolerated by the result clustering approach (see Chapter 4.6). This means, small green interval changes are in principle not detectable with the developed approach. Further research would be helpful that tries to overcome these restrictions. A potential solution could be the development of a sliding window-based approach which applies the developed methodology of this thesis for a fixed one hour time windows, whereas this window is slided over the time of day by small time steps, such as 5-10 mins. The described approach could be beneficial because it allows to obverse the signal timing parameters in a more continuous manner.

- **Improved self-verification:** Because trajectory volumes of the exploited commercially available FCD source were very small, the developed method required the aggregation and analysis of a long historical time period. The method assumes that the real signal planes are constant for every workday within the historical time period, which may not be always correct. Thus, it could happen for instance that on Mondays and Thursdays other signal plans are typical in use than on all other weekdays. If trajectory volumes are not dense enough to produce individual estimates for each workday, it may happen that the inferred estimates match only partially to the weekdays. A systematical trajectory analysis should be carried out that labels trajectories which are not matching with the derived signal timing estimates. These trajectories could be exploited to learn common features like, the weekday, the time of day and the month at which a very high number of trajectories are typically in contradiction with the inferred signal timing estimates. Such an approach would be

very helpful as it allows to detect systemically time periods in which trajectories do not fit the estimates. This information could be used in return to calculate better signal timing estimates for this particular time periods.

The research of this thesis underline the feasibility and thus the major relevance of a purely trajectory-based traffic signal state prediction even with nowadays commercially available trajectory volumes of a single FCD provider. The novel approaches and the comprehensive methodology of this thesis are an important contribution towards a network-wide and cost-efficient traffic signal state prediction. It may be assumed that the continuously increasing number of connected vehicles and smartphone devices will probably likewise cause a tremendous increase of vehicle trajectory data. This additional data would allow a more differentiated and accurate traffic signal state prediction.

APPENDIX A

Publications

This appendix lists author's publications that are related to this thesis.

Steffen Axer, Federico Pascucci, and Bernhard Friedrich. Estimation of traffic signal timing data and total delay for urban intersections based on low frequency floating car data. In *Mobil.Tum 2015*, München, Germany, 2015

Steffen Axer and Bernhard Friedrich. Estimating signal phase and timing for traffic actuated intersections based on low frequency floating car data. In *19th International Conference on Intelligent Transportation Systems*, Rio de Janeiro, 2016

Steffen Axer and Bernhard Friedrich. A methodology for signal timing estimation based on low frequency floating car data: Analysis of needed sample sizes and influencing factors. In *7th International Symposium on Highway Capacity and Quality of Service*, Berlin, Germany, 2016

Steffen Axer and Bernhard Friedrich. Signal timing estimation based on low frequency floating car data. In *14th World Conference on Transport Research*, Shanghai, China, 2016

Appendix B

Trajectory Volumes

B.1 Ingolstädter Straße / Frankfurter Ring

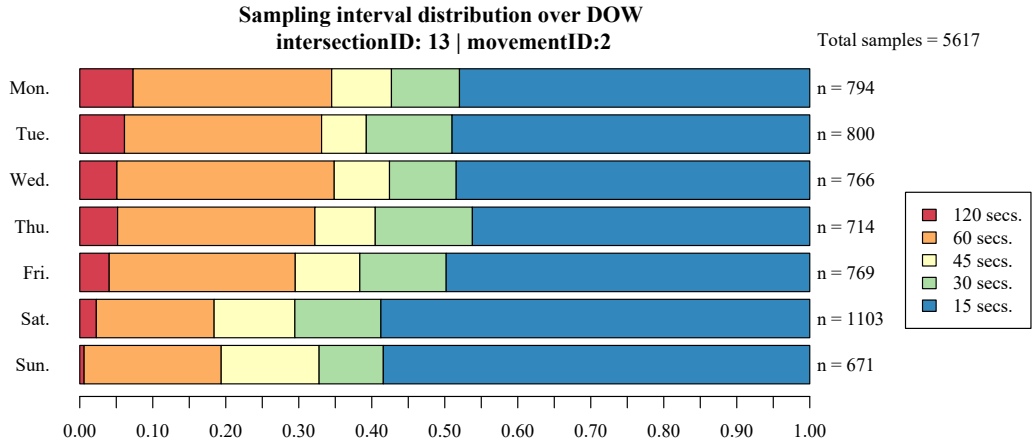


Figure B.1.1: Relative shares of different sampling intervals over DOW at Ingolstädterstraße / Frankfurter Ring, movement $m = 2$

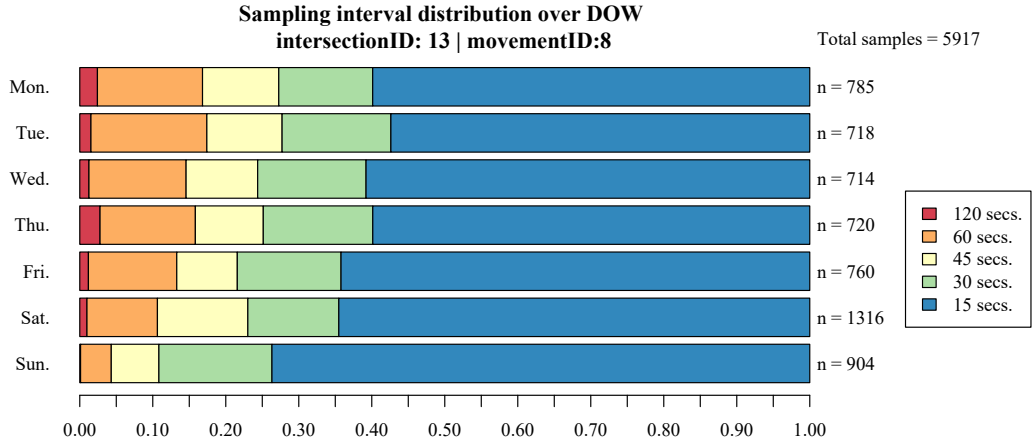


Figure B.1.2: Relative shares of different sampling intervals over DOW at Ingolstädterstraße / Frankfurter Ring, movement $m = 8$

B.2 Dachauer Straße / Georg-Brauchle-Ring

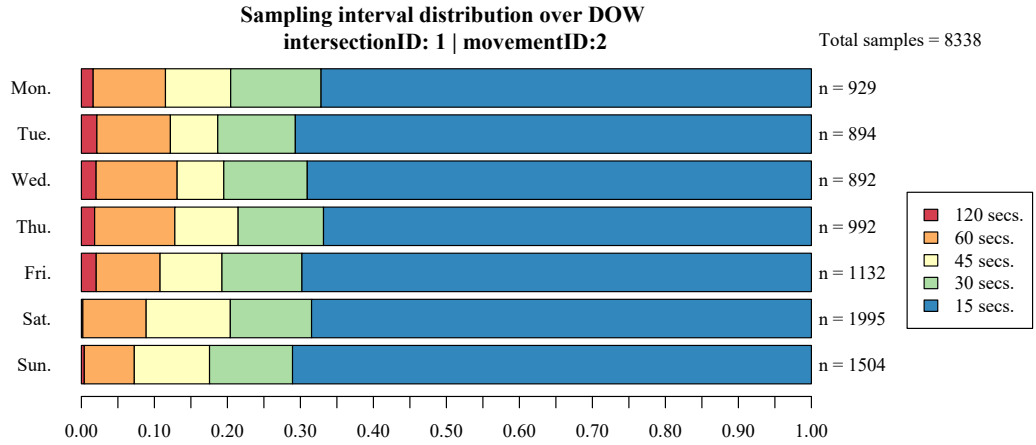


Figure B.2.3: Relative shares of different sampling intervals over DOW at Dachauer Straße / Georg-Brauchle-Ring, movement $m = 2$

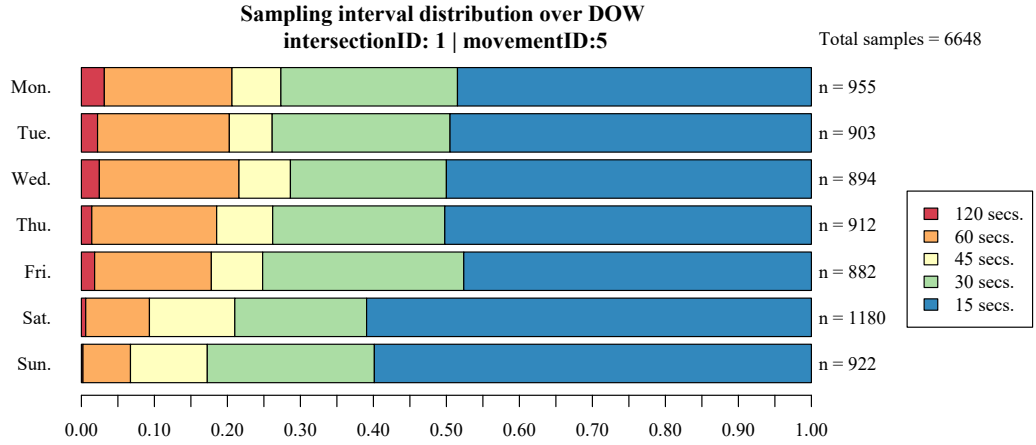


Figure B.2.4: Relative shares of different sampling intervals over DOW at Dachauer Straße / Georg-Brauchle-Ring, movement $m = 5$

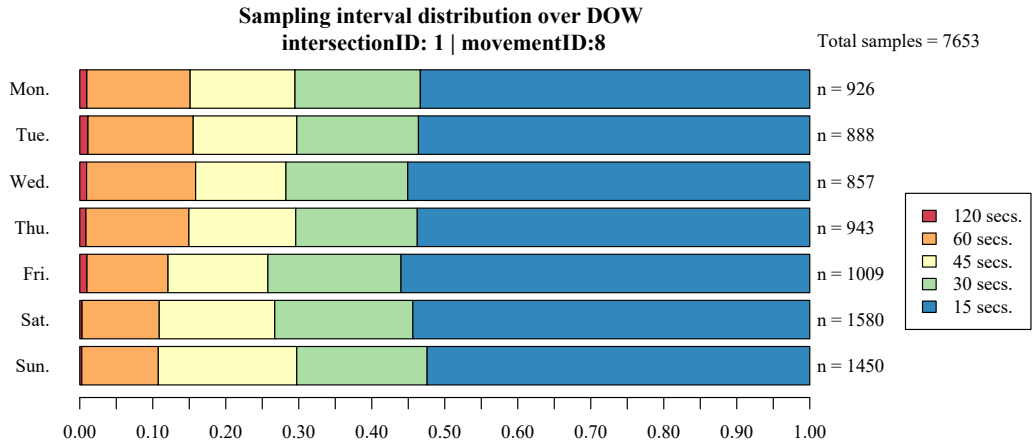


Figure B.2.5: Relative shares of different sampling intervals over DOW at Dachauer Straße / Georg-Brauchle-Ring, movement $m = 8$

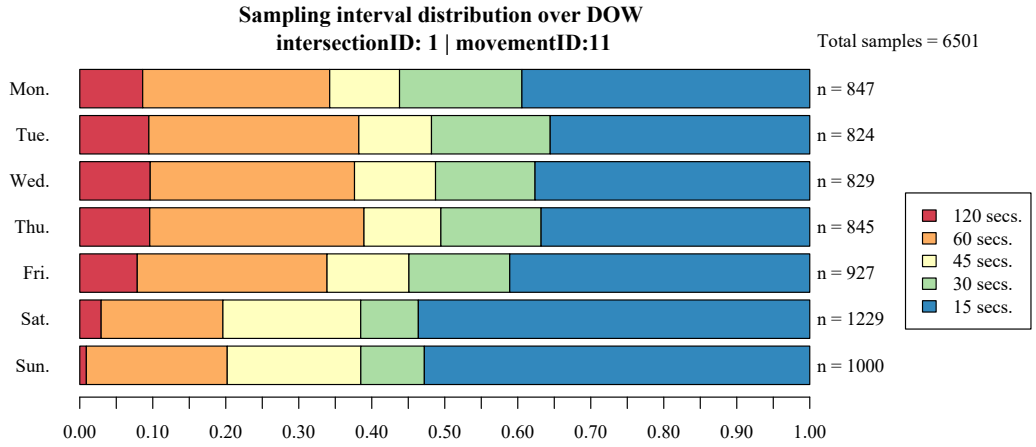


Figure B.2.6: Relative shares of different sampling intervals over DOW at Dachauer Straße / Georg-Brauchle-Ring, movement $m = 11$

APPENDIX C

Simulative Experiments

C.1 Estimation Output at Fixed-Time Intersection

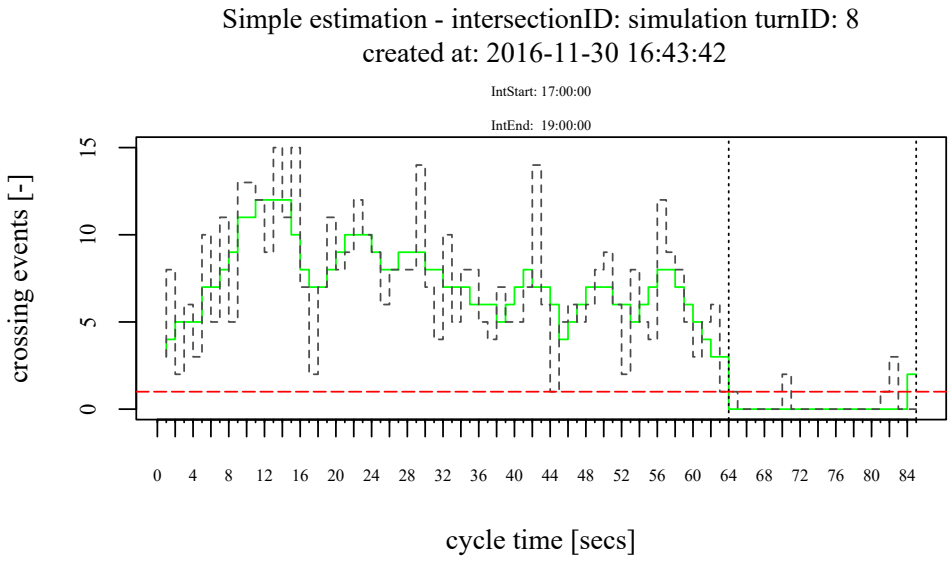


Figure C.1.1: Exemplary simple estimation result, fixed-time scenario ($x \approx 0.5$, $sample_{lvl} = 500$)

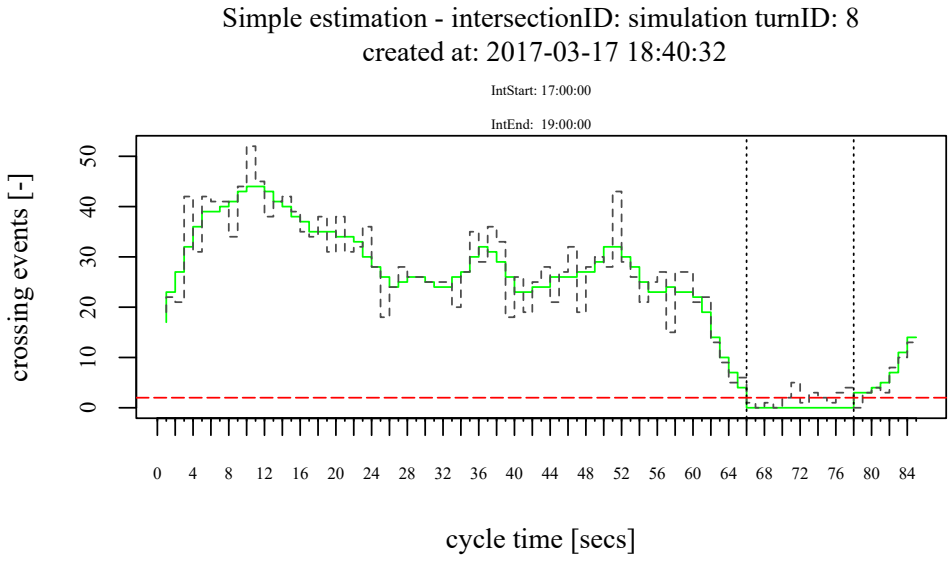


Figure C.1.2: Exemplary simple estimation result, fixed-time scenario ($x \approx 0.5$, $sample_{lvl} = 2000$)

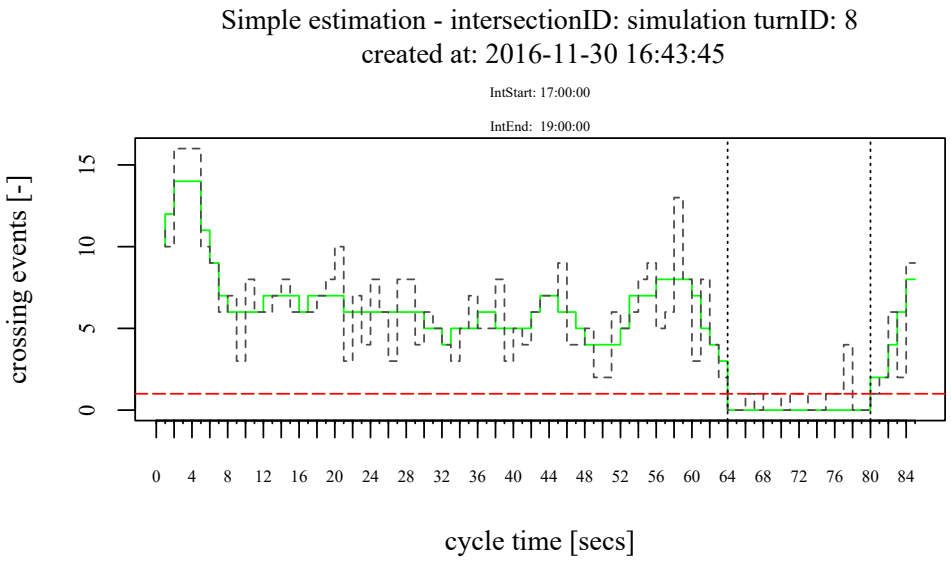


Figure C.1.3: Exemplary simple estimation result, fixed-time scenario ($x \approx 0.05$, $sample_{lvl} = 500$)

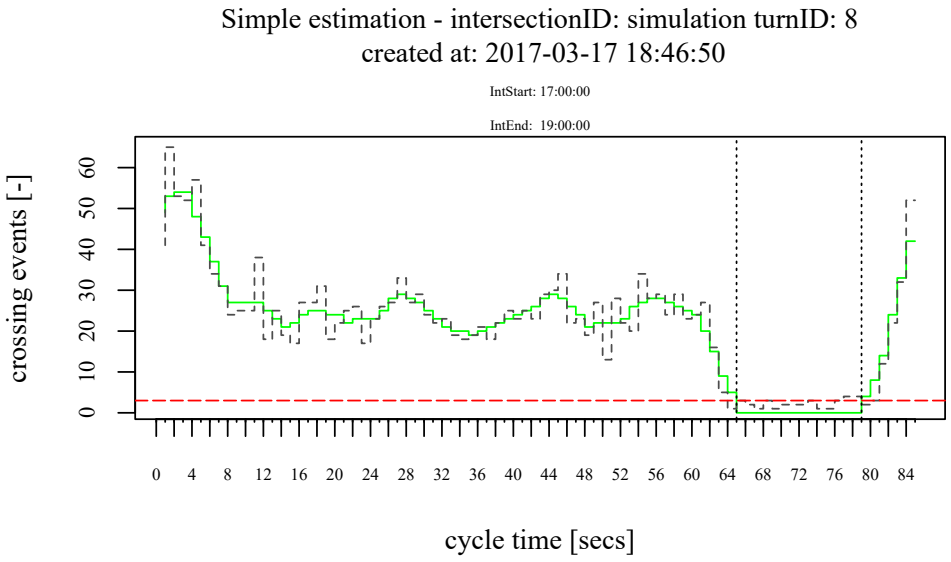


Figure C.1.4: Exemplary simple estimation result, fixed-time scenario ($x \approx 0.05$, $sample_{lvl} = 2000$)

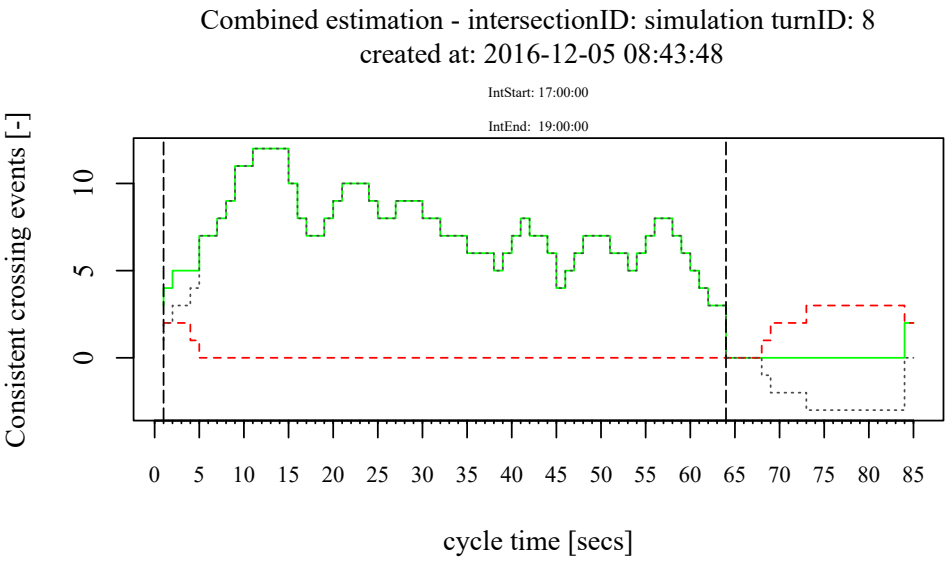


Figure C.1.5: Exemplary combined estimation result, fixed-time scenario ($x \approx 0.5$, $sample_{lvl} = 500$)

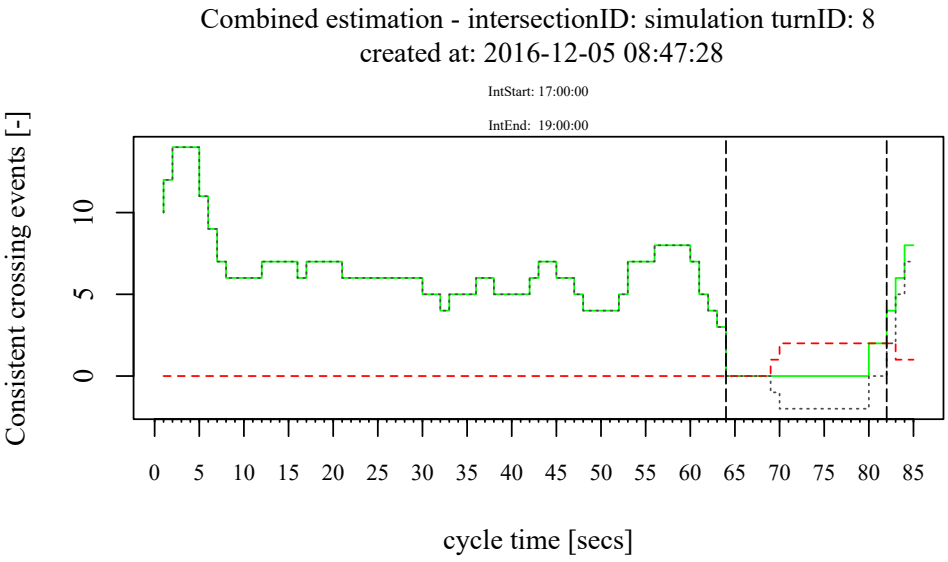


Figure C.1.6: Exemplary combined estimation result, fixed-time scenario ($x \approx 0.05$, $sample_{lvl} = 500$)

C.2 Estimation Output - Trajectory Filtering

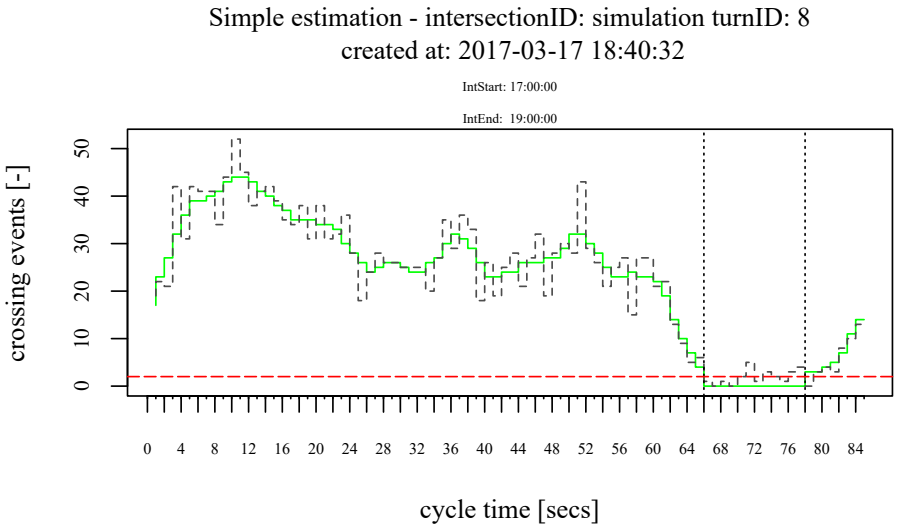


Figure C.2.7: Exemplary simple estimation result, fixed-time scenario ($x \approx 0.5$, $sample_{lvl} = 2000$, **enabled trajectory filtering**)

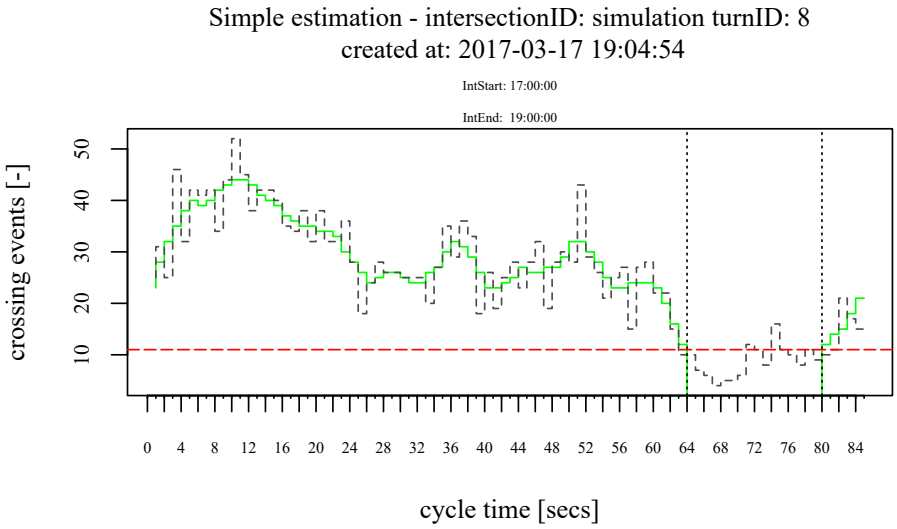


Figure C.2.8: Exemplary simple estimation result, fixed-time scenario ($x \approx 0.5$, $sample_{lvl} = 2000$, **disabled trajectory filtering**)

C.3 Estimation Output at Actuated Intersection

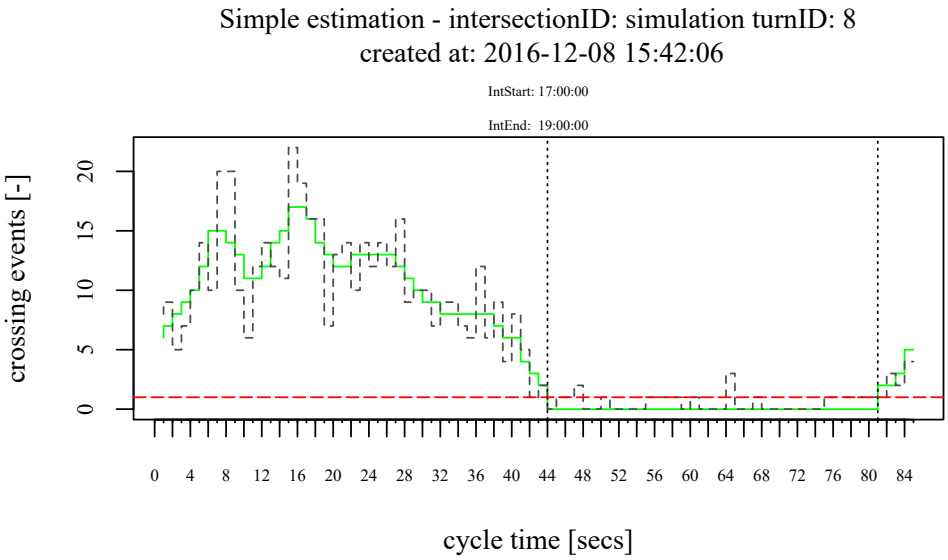


Figure C.3.9: Exemplary simple estimation result, traffic actuated scenario ($x \approx 1$, $sample_{lvl} = 500$)

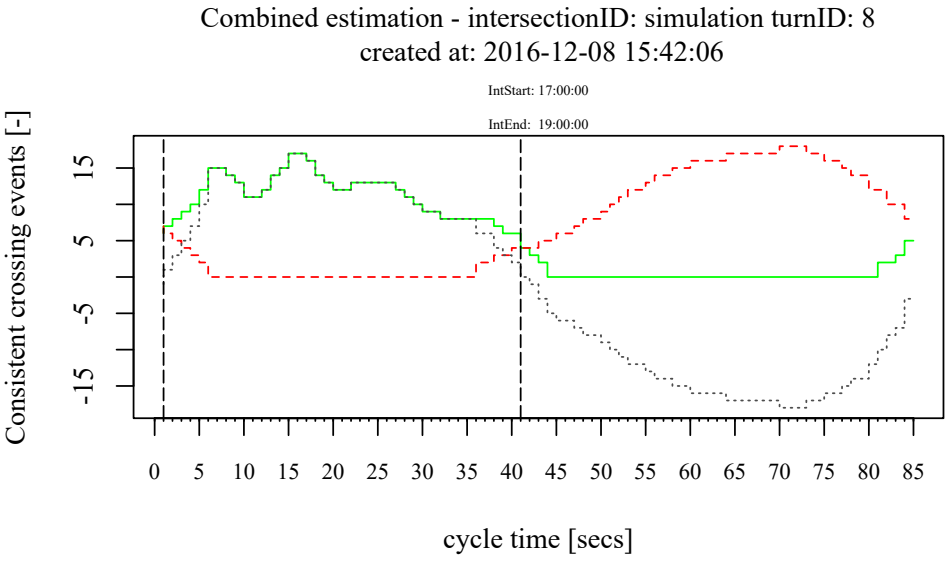


Figure C.3.10: Exemplary combined estimation result, traffic actuated scenario ($x \approx 1$, $sample_{lvl} = 500$)

C.4 Exemplary Clustering and Time Assignment

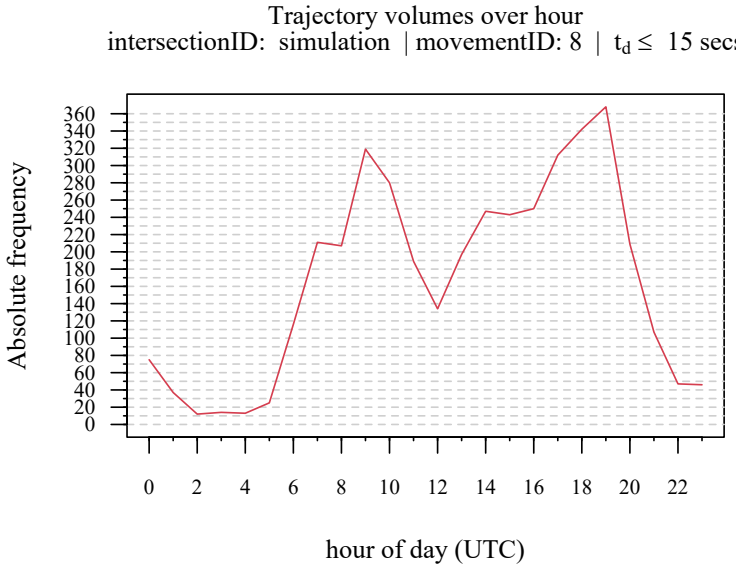


Figure C.4.11: Trajectory volumes of a simulated day, used to verify clustering and time assignment algorithm

Movement = 8						Movement = 2								
h	**C**	**tg_{start}**	**tg_{end}**	**$tg_{start,si.}$**	**$tg_{end,si.}$**	**h**	**C**	**tg_{start}**	**tg_{end}**	**$tg_{start,si.}$**	**$tg_{end,si.}$**	**cls_{best}**	**$TA_{i=2}$**	**$TA_{i=3}$**
6	60	60	34	58	34	6	60	1	33	57	33	1_8_60_1	6	6
7	85	1	49	81	50	7	85	1	50	84	50	2_8_85_1	7	7
8	85	1	49	81	50	8	85	1	50	84	50	2_8_85_1	8	8
9	85	1	49	81	50	9	85	1	50	84	50	2_8_85_1	9	9
10	60	60	34	58	34	10	60	1	33	57	33	1_8_60_1	10	10
11	60	60	34	58	34	11	60	1	33	57	33	1_8_60_1	11	11
12	60	60	34	58	34	12	60	1	33	57	33	1_8_60_1	12	12
13	60	60	34	58	34	13	60	1	33	57	33	1_8_60_1	13	13
14	60	60	34	58	34	14	60	1	33	57	33	1_8_60_1	14	14
15	85	1	49	81	50	15	85	1	50	84	50	2_8_85_1	15	15
16	85	1	49	81	50	16	85	1	50	84	50	2_8_85_1	16	16
17	85	1	49	81	50	17	85	1	50	84	50	2_8_85_1	17	17
18	85	1	49	81	50	18	85	1	50	84	50	2_8_85_1	18	18
19	45	44	24	41	25	19	45	45	23	42	23	3_8_45_1	19	19
20	45	44	24	41	25	20	45	45	23	42	23	3_8_45_1	20	20
21	45	44	24	41	25	21	45	45	23	42	23	3_8_45_1	21	21

Figure C.4.12: Estimation results after second update of time assignment table

Appendix D

Real World Experiments

D.1 Ingolstädter Straße / Frankfurter Ring

D.1.1 Estimated Green Intervals

Table D.1: Estimation results at intersection Ingolstädter Straße / Frankfurter Ring, id=13, backward calculation method = RRV

cluster	status	m	$class$	C [secs]	tg_{start} [secs]	tg_{end} [secs]
2_2_90_2	not adjusted	1	1	90	88	23
2_2_90_2	adjusted	2	3	90	76	26
2_2_90_2	adjusted	8	3	90	84	11
2_2_90_2	adjusted	10	3	90	33	51
2_2_90_2	adjusted	12	3	90	33	52
1_2_90_3	not adjusted	1	1	90	59	29
1_2_90_3	adjusted	2	3	90	60	89
1_2_90_3	adjusted	4	1	90	36	53
1_2_90_3	adjusted	6	3	90	31	50
1_2_90_3	adjusted	8	3	90	54	3
1_2_90_3	adjusted	9	3	90	57	72
1_2_90_3	adjusted	10	3	90	7	25
1_2_90_3	adjusted	12	3	90	7	30
4_2_90_1	not adjusted	2	1	90	84	20
3_2_90_2	not adjusted	2	1	90	53	87
3_2_90_2	not adjusted	8	1	90	52	76

Table D.2: Time assignment table at intersection Ingolstädter Straße / Frankfurter Ring, id=13

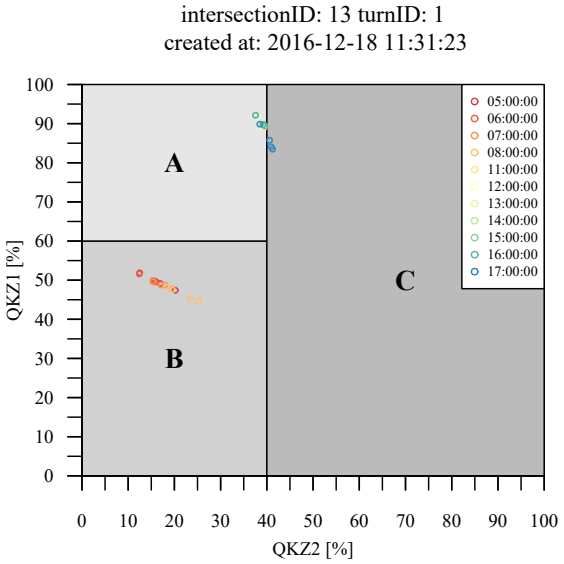
cluster	h
4_2_90_1	04:00:00
2_2_90_2	05:00:00
2_2_90_2	06:00:00
2_2_90_2	07:00:00
2_2_90_2	08:00:00
1_2_90_3	11:00:00
1_2_90_3	12:00:00
1_2_90_3	13:00:00
1_2_90_3	14:00:00
1_2_90_3	15:00:00
1_2_90_3	16:00:00
1_2_90_3	17:00:00
3_2_90_2	18:00:00

D.1.2 Available Trajectory Volumes

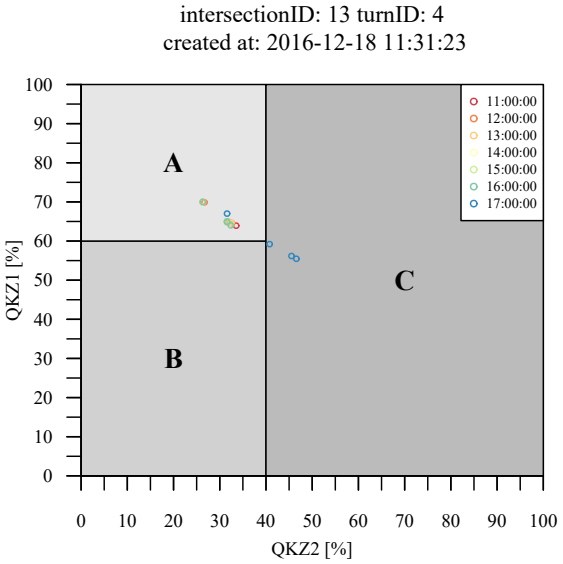
Table D.3: Trajectory volumes [veh/h] at intersection Ingolstädter Straße / Frankfurter Ring, id=13, $t_d \leq 15$ secs

	Movement											
h	1	2	3	4	5	6	7	8	9	10	11	12
0	29	58	9	4	-	11	9	44	17	28	-	27
1	31	64	4	9	-	6	6	65	21	34	-	26
2	19	53	5	5	-	14	7	37	16	29	-	17
3	33	68	15	19	-	2	1	39	31	49	-	31
4	67	209	40	61	-	23	31	88	43	63	-	72
5	147	243	53	84	-	41	36	107	61	100	-	126
6	157	267	63	65	-	45	36	162	75	115	-	119
7	197	270	50	63	-	44	39	156	88	125	-	122
8	215	264	44	64	-	52	63	163	91	136	-	195
9	258	227	41	59	-	39	64	231	135	167	-	210
10	369	239	60	77	-	59	55	211	123	151	-	222
11	353	275	56	72	-	43	65	230	140	167	-	258
12	297	317	40	63	-	69	52	255	133	156	-	219
13	287	333	58	65	-	59	72	250	108	148	-	223
14	257	355	51	65	-	75	60	234	98	165	-	180
15	310	341	47	62	-	77	57	279	122	146	-	199
16	439	510	79	71	-	117	68	332	158	229	-	288
17	456	551	72	84	-	108	67	303	190	266	-	286
18	325	336	46	50	-	65	38	219	153	207	-	180
19	201	206	28	20	-	54	43	156	111	134	-	122
20	143	171	21	15	-	36	31	134	90	135	-	118
21	119	136	17	21	-	34	21	137	74	84	-	77
22	98	96	14	16	-	26	21	88	48	76	-	65
23	42	78	7	12	-	8	8	71	36	49	-	39

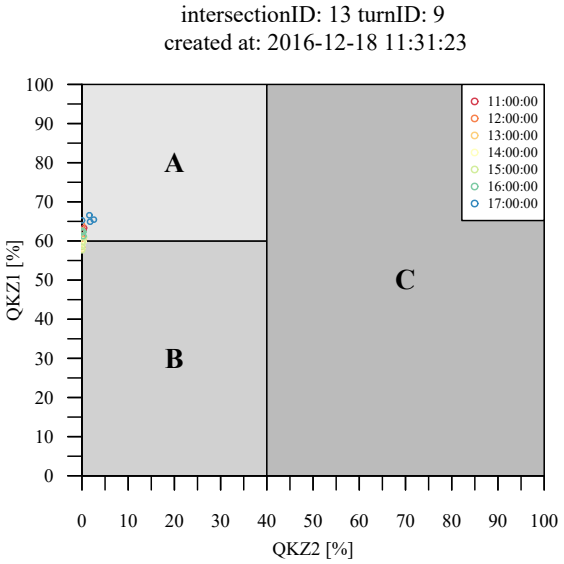
D.1.3 Estimation Quality



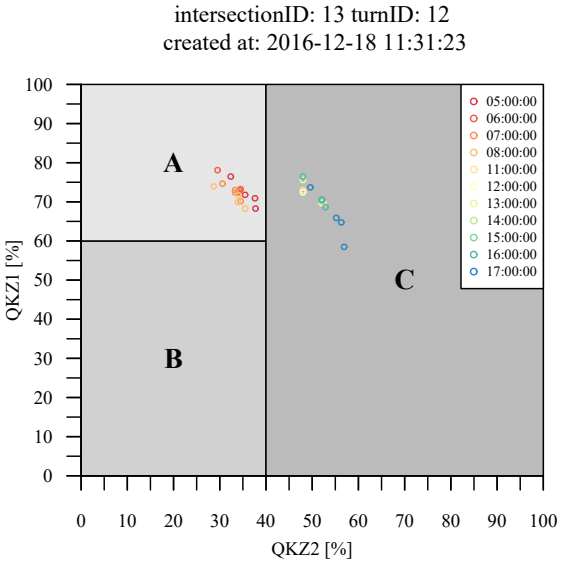
(a) QKZ for $m = 1$



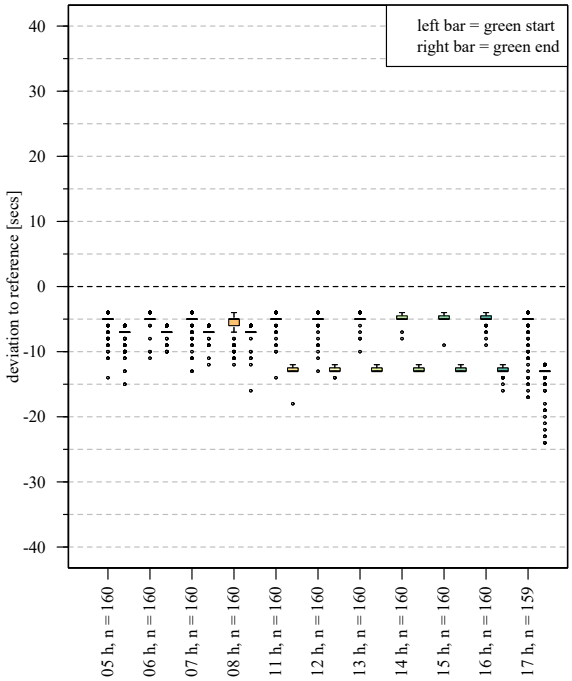
(b) QKZ for $m = 4$



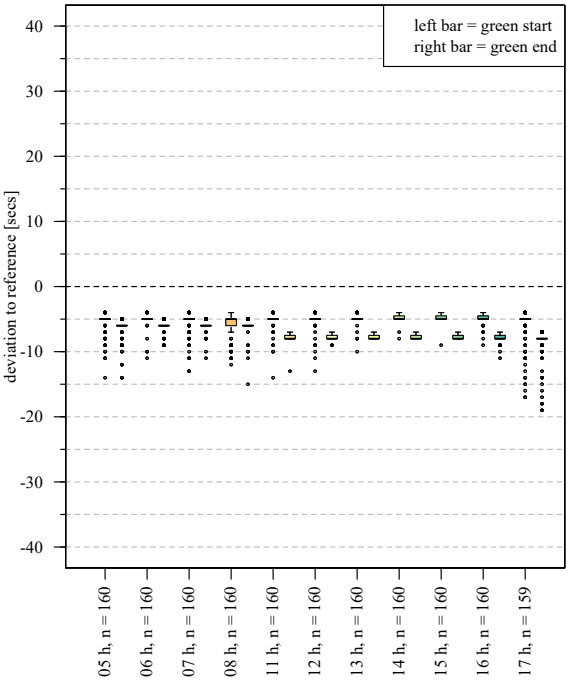
(c) QKZ for $m = 9$



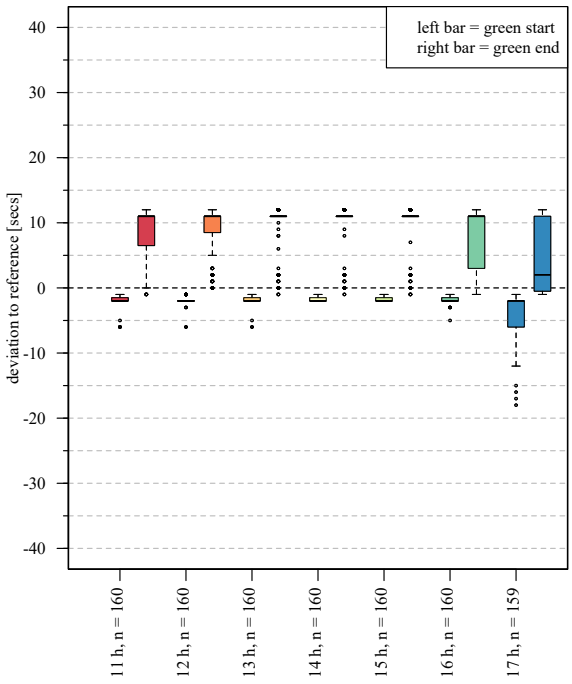
(d) QKZ for $m = 12$



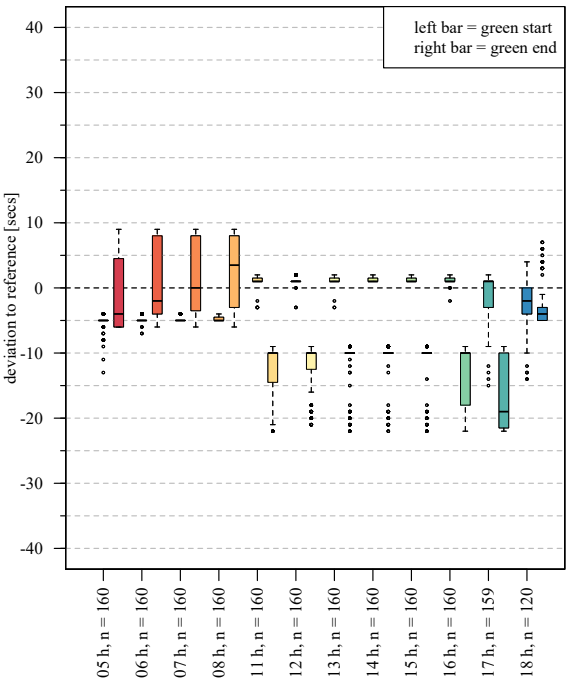
(a) Green accuracy for $m = 12$



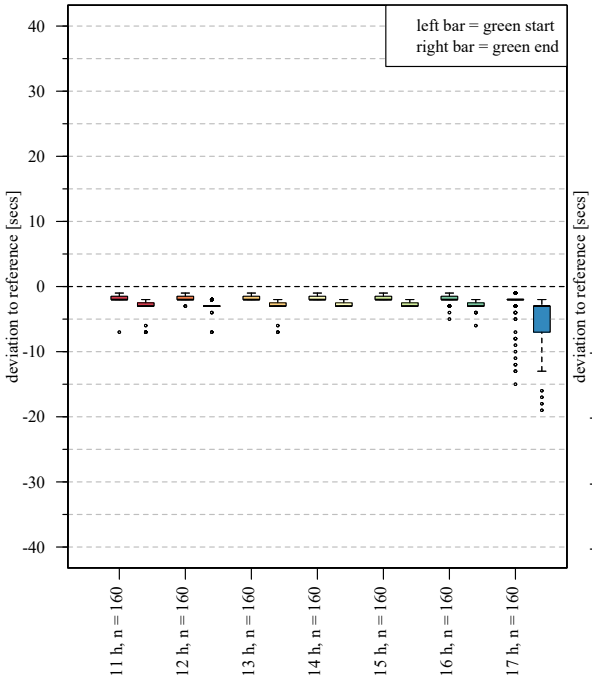
(b) Green accuracy for for $m = 10$



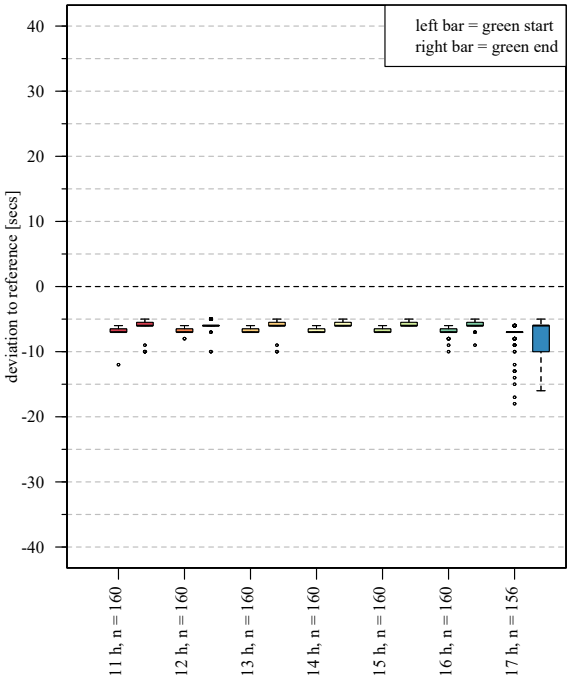
(c) Green accuracy for $m = 9$



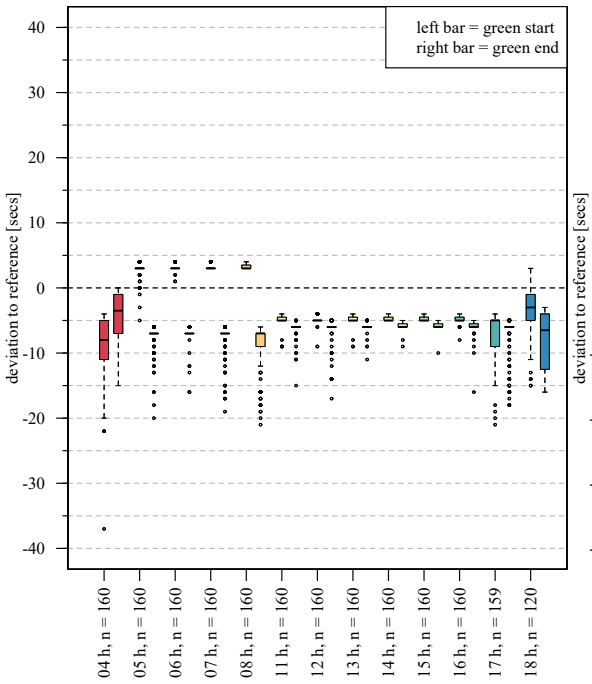
(d) Green accuracy for $m = 8$



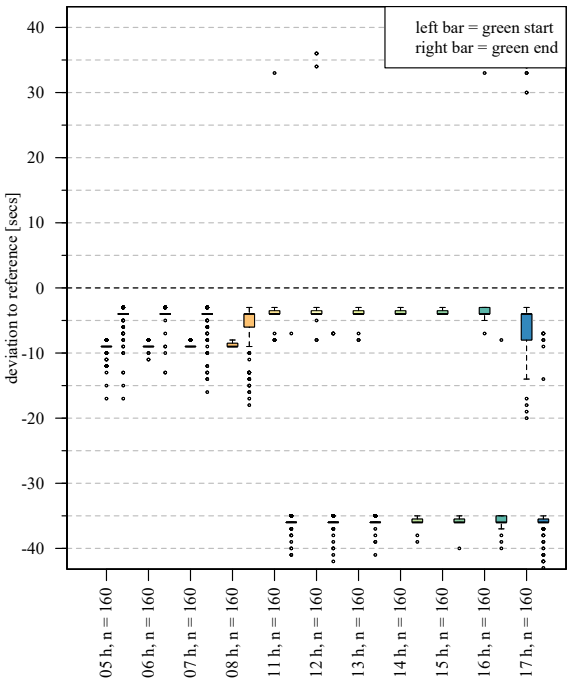
(a) Green accuracy for $m = 6$



(b) Green accuracy for for $m = 4$



(c) Green accuracy for $m = 2$



(d) Green accuracy for $m = 1$

D.2 Ingolstädter Straße / Neuherbergstraße

D.2.1 Estimated Green Intervals

Table D.4: Estimation results at intersection Ingolstädter Straße / Neuherbergstraße, id=2, backward calculation method = RRV

cluster	status	m	$class$	C [secs]	tg_{start} [secs]	tg_{end} [secs]
3_2_90_1	not adjusted	2	1	90	8	45
2_2_90_2	adjusted	2	3	90	1	49
2_2_90_2	adjusted	8	3	90	3	55
2_2_90_2	adjusted	12	3	90	64	86
2_2_90_2	not adjusted	1	1	90	4	47
1_2_90_3	adjusted	2	3	90	46	3
1_2_90_3	adjusted	8	3	90	45	10
1_2_90_3	adjusted	12	3	90	16	38
1_2_90_3	not adjusted	1	1	90	52	2
4_8_90_1	not adjusted	8	1	90	54	71

Table D.5: Time assignment table at intersection Ingolstädter Straße / Neuherbergstraße, id=2

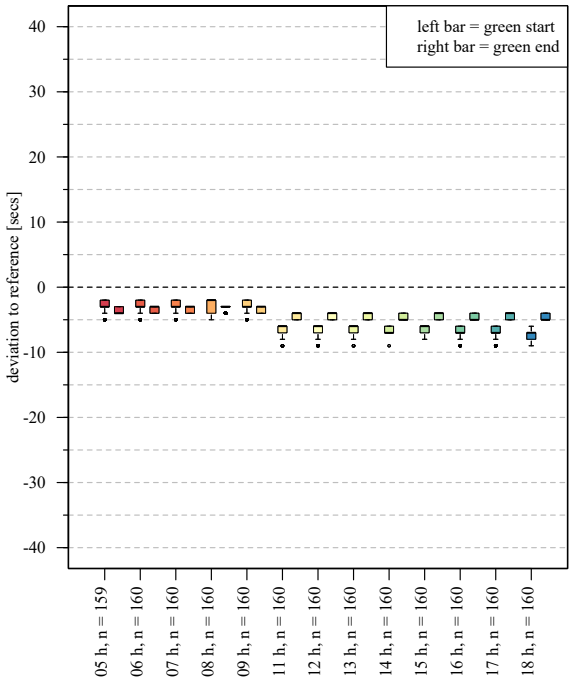
cluster	h
3_2_90_1	04:00:00
2_2_90_2	05:00:00
2_2_90_2	06:00:00
2_2_90_2	07:00:00
2_2_90_2	08:00:00
2_2_90_2	09:00:00
4_8_90_1	10:00:00
1_2_90_3	11:00:00
1_2_90_3	12:00:00
1_2_90_3	13:00:00
1_2_90_3	14:00:00
1_2_90_3	15:00:00
1_2_90_3	16:00:00
1_2_90_3	17:00:00
1_2_90_3	18:00:00

D.2.2 Available Trajectory Volumes

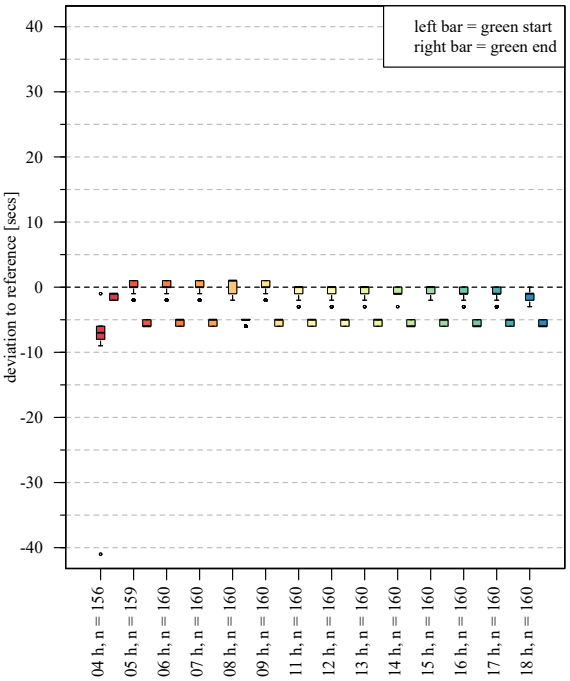
Table D.6: Trajectory volumes [veh/h] at intersection Ingolstädter Straße / Neuherbergstraße, id=1, $t_d \leq 15$ secs

	Movement											
h	1	2	3	4	5	6	7	8	9	10	11	12
0	11	25	-	-	-	-	-	30	10	4	-	31
1	25	25	-	-	-	-	-	34	9	5	-	15
2	12	17	-	-	-	-	-	27	4	6	-	33
3	37	67	-	-	-	-	-	24	7	7	-	41
4	134	208	-	-	-	-	-	69	11	21	-	80
5	273	480	-	-	-	-	-	135	29	24	-	113
6	306	480	-	-	-	-	-	200	25	24	-	183
7	160	378	-	-	-	-	-	237	18	22	-	248
8	147	385	-	-	-	-	-	309	12	30	-	249
9	135	312	-	-	-	-	-	344	28	28	-	253
10	141	377	-	-	-	-	-	395	32	26	-	209
11	154	317	-	-	-	-	-	425	25	38	-	255
12	162	364	-	-	-	-	-	435	31	32	-	269
13	161	345	-	-	-	-	-	426	24	23	-	219
14	205	354	-	-	-	-	-	447	29	22	-	229
15	222	423	-	-	-	-	-	516	32	25	-	281
16	273	584	-	-	-	-	-	577	39	34	-	329
17	252	547	-	-	-	-	-	538	33	53	-	288
18	276	349	-	-	-	-	-	371	36	32	-	268
19	253	222	-	-	-	-	-	169	32	27	-	159
20	126	167	-	-	-	-	-	147	11	14	-	77
21	69	130	-	-	-	-	-	146	26	16	-	58
22	46	89	-	-	-	-	-	98	20	8	-	52
23	25	44	-	-	-	-	-	59	3	6	-	20

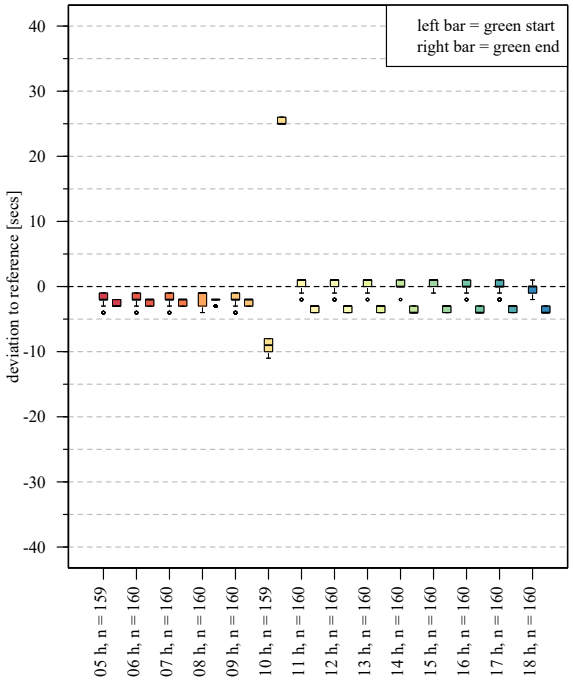
D.2.3 Estimation Quality



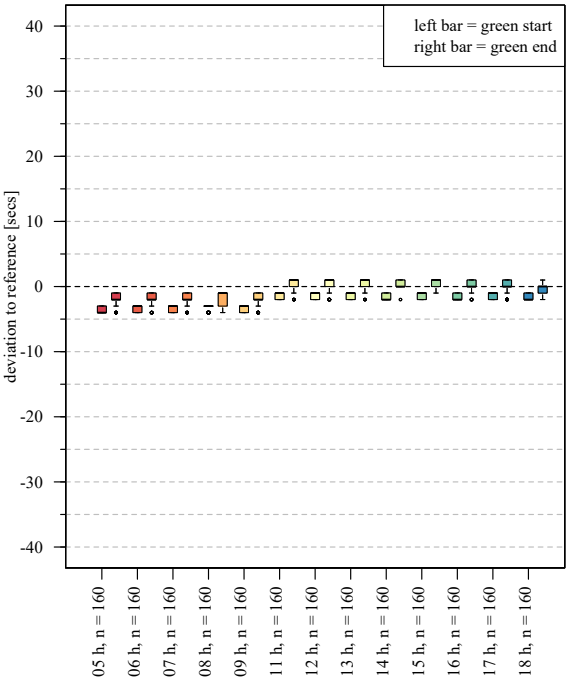
(a) Green accuracy for $m = 1$



(b) Green accuracy for for $m = 2$



(c) Green accuracy for $m = 8$



(d) Green accuracy for $m = 12$

D.3 Dachauer Straße / Georg-Brauchle-Ring

D.3.1 Estimated Green Intervals

Table D.7: Estimation results at intersection Dachauer Straße / Georg-Brauchle-Ring, id=1, backward calculation method = RRV

cluster	status	m	$class$	C [secs]	tg_{start} [secs]	tg_{end} [secs]	p_i
7_2_90_1	not adjusted	2	1	90	75	10	0.0
4_2_90_2	adjusted	2	3	90	78	27	0.4
4_2_90_2	adjusted	5	3	90	35	58	0.4
4_2_90_2	adjusted	8	3	90	82	10	0.4
4_2_90_2	adjusted	11	3	90	36	74	0.4
6_2_90_2	adjusted	2	3	90	71	17	0.2
6_2_90_2	adjusted	5	3	90	31	51	0.2
3_8_90_2	adjusted	2	3	90	64	22	0.4
3_8_90_2	adjusted	5	3	90	25	56	0.4
3_8_90_2	adjusted	8	3	90	67	12	0.4
3_8_90_2	adjusted	11	3	90	31	58	0.4
5_2_90_3	adjusted	2	3	90	64	90	0.2
5_2_90_3	adjusted	5	3	90	29	55	0.4
5_2_90_3	adjusted	8	3	90	67	4	0.2
1_8_90_4	adjusted	2	3	90	67	18	0.4
1_8_90_4	adjusted	3	3	90	6	21	0.5
1_8_90_4	adjusted	5	3	90	30	59	0.8
1_8_90_4	adjusted	8	3	90	67	18	0.4
1_8_90_4	adjusted	9	3	90	3	21	0.5
1_8_90_4	adjusted	11	3	90	30	59	0.8
2_2_90_6	adjusted	2	3	90	59	2	0.4
2_2_90_6	adjusted	3	3	90	4	14	0.5
2_2_90_6	adjusted	5	3	90	21	51	0.6
2_2_90_6	adjusted	8	3	90	64	1	0.4
2_2_90_6	adjusted	11	3	90	23	55	0.6

Table D.8: Time assignment table at intersection Dachauer Straße / Georg-Brauchle-Ring, id=1

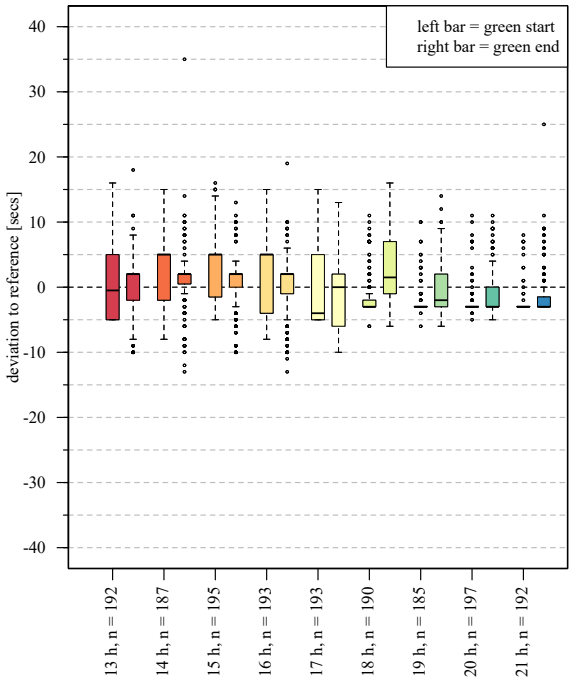
cluster	h
7_2_90_1	04:00:00
4_2_90_2	05:00:00
4_2_90_2	06:00:00
6_2_90_2	07:00:00
3_8_90_2	08:00:00
3_8_90_2	09:00:00
3_8_90_2	10:00:00
3_8_90_2	11:00:00
5_2_90_3	12:00:00
1_8_90_4	13:00:00
1_8_90_4	14:00:00
1_8_90_4	15:00:00
1_8_90_4	16:00:00
1_8_90_4	17:00:00
2_2_90_6	18:00:00
2_2_90_6	19:00:00
2_2_90_6	20:00:00
2_2_90_6	21:00:00

D.3.2 Available Trajectory Volumes

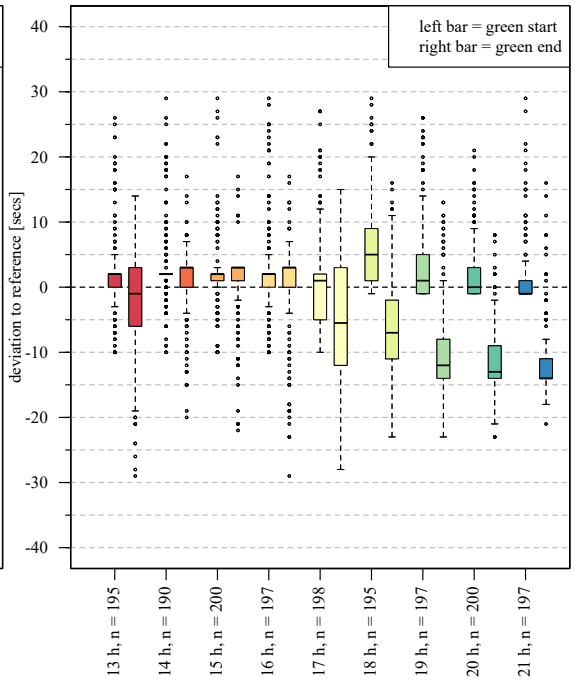
Table D.9: Trajectory volumes [veh/h] at intersection Dachauer Straße / Georg-Brauchle-Ring, id=1, $t_d \leq 15$ secs

	Movement											
h	1	2	3	4	5	6	7	8	9	10	11	12
0	-	66	12	-	60	-	-	98	30	-	36	7
1	-	45	12	-	84	-	-	66	21	-	36	10
2	-	68	26	-	37	-	-	67	20	-	32	10
3	-	106	17	-	42	-	-	63	21	-	55	13
4	-	185	50	-	103	-	-	123	28	-	152	65
5	-	329	67	-	117	-	-	168	80	-	318	76
6	-	400	94	-	194	-	-	270	76	-	383	66
7	-	463	93	-	286	-	-	302	93	-	397	72
8	-	522	114	-	308	-	-	360	121	-	382	61
9	-	474	88	-	363	-	-	413	136	-	340	73
10	-	527	90	-	395	-	-	475	145	-	396	91
11	-	537	93	-	393	-	-	475	160	-	399	101
12	-	543	111	-	438	-	-	492	168	-	419	102
13	-	527	83	-	400	-	-	543	171	-	390	111
14	-	559	82	-	465	-	-	588	167	-	382	97
15	-	571	94	-	577	-	-	651	232	-	394	105
16	-	776	129	-	774	-	-	804	323	-	499	137
17	-	827	138	-	790	-	-	761	250	-	488	143
18	-	683	172	-	392	-	-	575	180	-	331	111
19	-	448	141	-	255	-	-	450	111	-	238	98
20	-	321	60	-	170	-	-	366	104	-	194	63
21	-	266	34	-	146	-	-	342	86	-	177	54
22	-	182	27	-	121	-	-	238	81	-	101	49
23	-	94	18	-	91	-	-	151	42	-	68	25

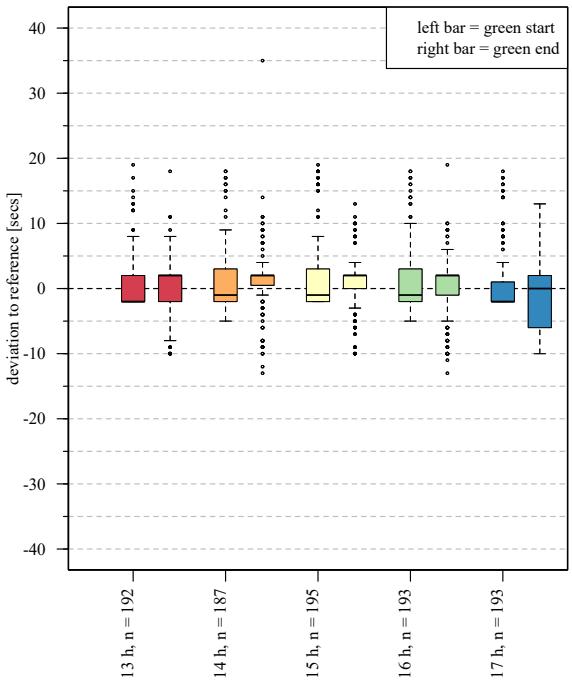
D.3.3 Estimation Quality



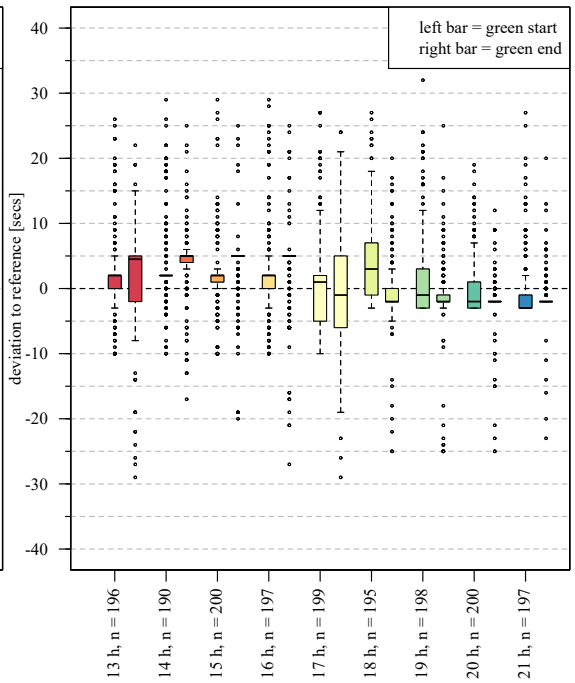
(a) Green accuracy for $m = 3$



(b) Green accuracy for for $m = 5$

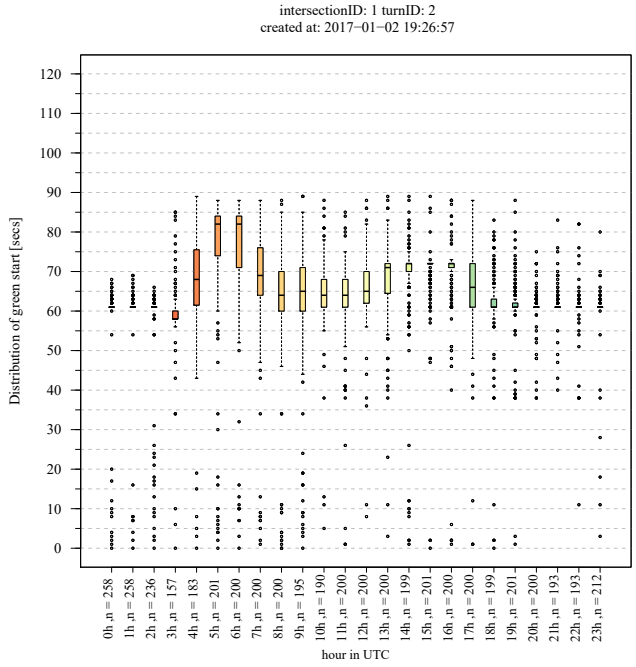


(c) Green accuracy for $m = 9$

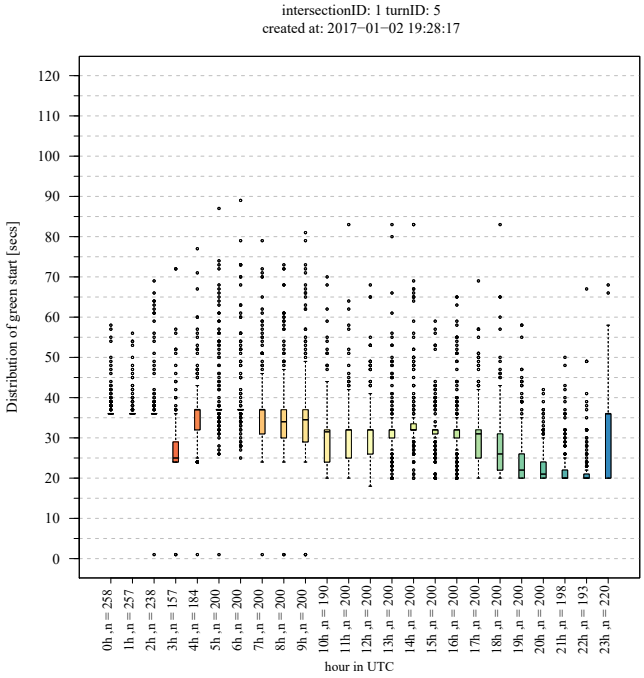


(d) Green accuracy for $m = 11$

D.3.4 Ground Truth Signal Timing

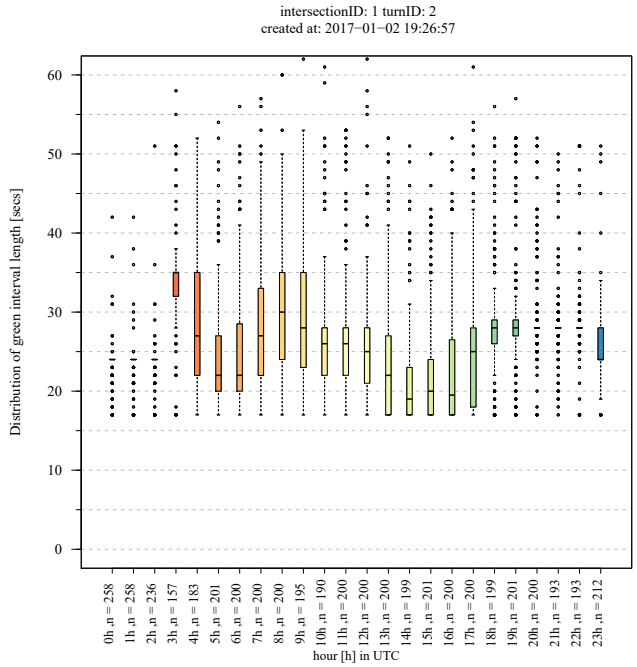


(a) Movement $m = 2$

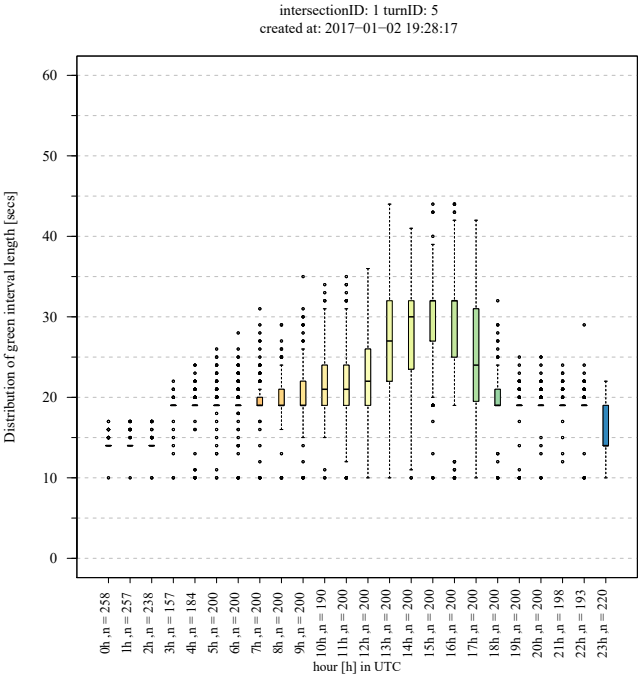


(b) Movement $m = 5$

Figure D.3.6: Distribution of green starts



(a) Movement $m = 2$



(b) Movement $m = 5$

Figure D.3.7: Distribution of green interval length

D.4 Frankfurter Ring / Ungererstraße

D.4.1 Estimated Green Intervals

Table D.10: Estimation results at intersection Frankfurter Ring / Ungererstraße, id=16, backward calculation method = RRV

cluster	status	m	$class$	C [secs]	tg_{start} [secs]	tg_{end} [secs]	p_i
2_12_90_2	adjusted	5	3	90	26	54	0.6
2_12_90_2	adjusted	11	3	90	48	75	0.4
2_12_90_2	adjusted	8	3	90	84	10	0.6
2_12_90_2	adjusted	2	3	90	81	21	0.6
2_12_90_2	adjusted	12	3	90	34	77	0.5
2_12_90_2	adjusted	1	3	90	84	19	1.0
1_8_90_2	adjusted	5	3	90	7	32	0.6
1_8_90_2	adjusted	11	3	90	25	61	0.4
1_8_90_2	adjusted	8	3	90	70	1	0.8
1_8_90_2	adjusted	2	3	90	68	1	0.8
1_8_90_2	adjusted	12	3	90	31	63	0.5
1_8_90_2	adjusted	6	3	90	9	28	0.5
1_8_90_2	adjusted	1	3	90	67	2	1.0
4_12_90_1	not adjusted	12	1	90	53	73	0.0
5_12_90_1	not adjusted	12	1	90	37	71	0.0
3_12_90_3	not adjusted	12	1	90	30	57	0.0
6_12_70_1	not adjusted	12	1	70	8	26	0.0

Table D.11: Time assignment table at intersection Frankfurter Ring / Ungererstraße, id=16

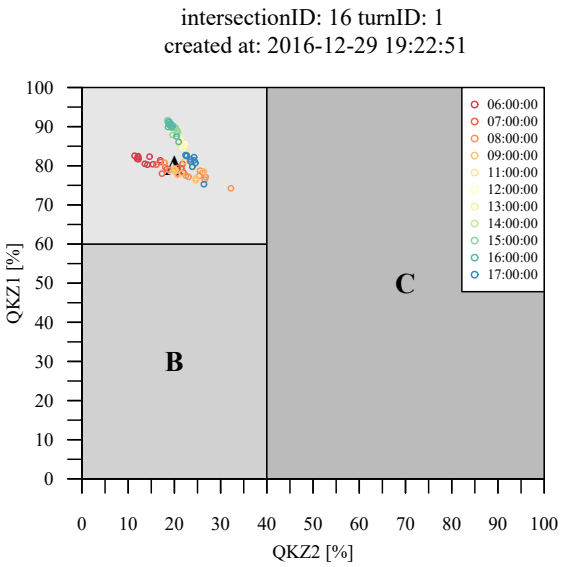
cluster	h
4_12_90_1	05:00:00
2_12_90_2	06:00:00
2_12_90_2	07:00:00
2_12_90_2	08:00:00
2_12_90_2	09:00:00
5_12_90_1	10:00:00
1_8_90_2	11:00:00
1_8_90_2	12:00:00
1_8_90_2	13:00:00
1_8_90_2	14:00:00
1_8_90_2	15:00:00
1_8_90_2	16:00:00
1_8_90_2	17:00:00
3_12_90_3	18:00:00
6_12_70_1	19:00:00

D.4.2 Available Trajectory Volumes

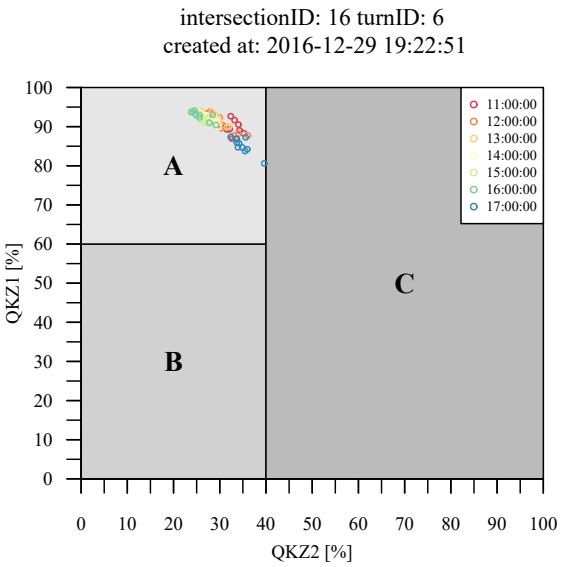
Table D.12: Trajectory volumes [veh/h] at intersection Frankfurter Ring / Ungererstraße, id=16, $t_d \leq 15$ secs

	Movement											
h	1	2	3	4	5	6	7	8	9	10	11	12
0	10	23	4	5	19	13	19	27	18	-	18	59
1	5	20	7	3	12	7	22	27	20	-	23	57
2	9	20	11	2	8	4	17	33	9	-	14	58
3	10	28	7	2	13	4	30	406	7	-	25	74
4	43	48	15	5	37	17	39	736	14	-	57	152
5	68	127	28	14	147	46	69	874	31	-	81	202
6	91	191	34	18	192	68	144	913	43	-	131	241
7	54	186	37	38	229	48	126	929	65	-	120	317
8	69	144	55	38	195	63	140	1043	68	-	142	418
9	49	114	45	26	166	62	154	968	54	-	167	409
10	71	123	40	37	179	54	156	1016	60	-	180	384
11	59	111	43	32	161	64	146	1054	62	-	167	500
12	53	108	43	31	173	52	135	1062	65	-	185	511
13	48	117	43	28	185	55	141	1053	57	-	176	496
14	72	118	31	34	183	78	141	1030	53	-	167	526
15	61	163	45	37	231	108	183	1141	53	-	200	590
16	102	244	67	54	254	131	243	1122	84	-	264	647
17	95	279	61	39	262	146	211	1229	79	-	229	574
18	74	184	40	23	196	85	151	804	61	-	151	412
19	57	108	27	16	148	60	95	251	56	-	103	337
20	43	88	24	17	121	39	97	111	45	-	98	227
21	27	66	20	7	85	33	65	91	43	-	65	229
22	16	47	17	6	67	15	41	48	25	-	48	151
23	8	46	5	4	36	17	40	41	15	-	38	55

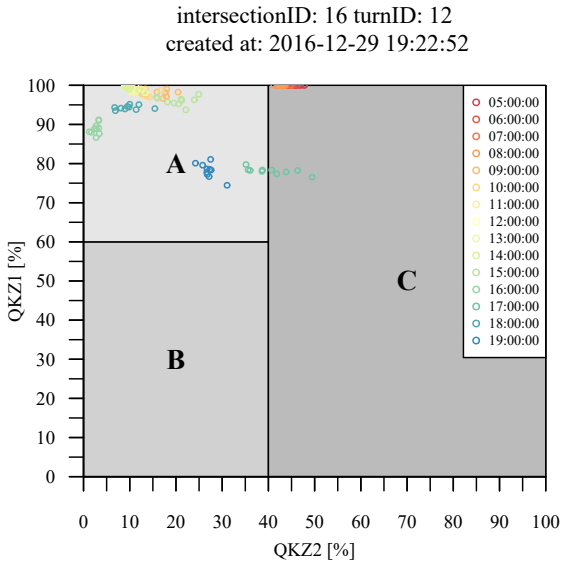
D.4.3 Estimation Quality



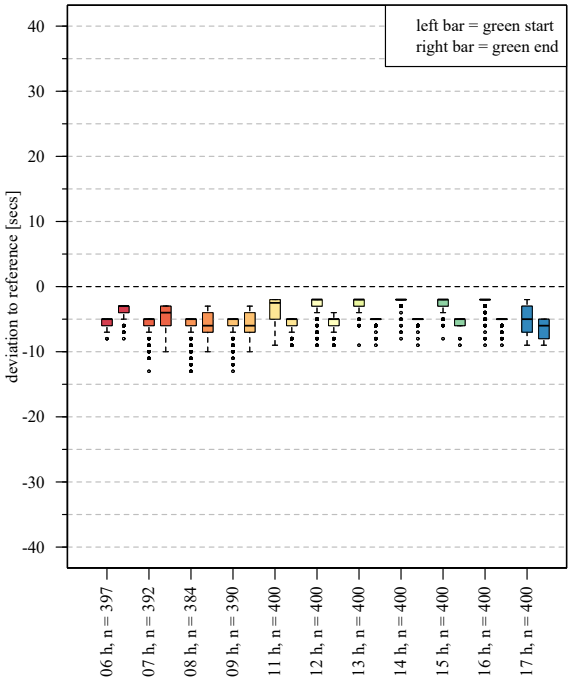
(a) QKZ for $m = 1$



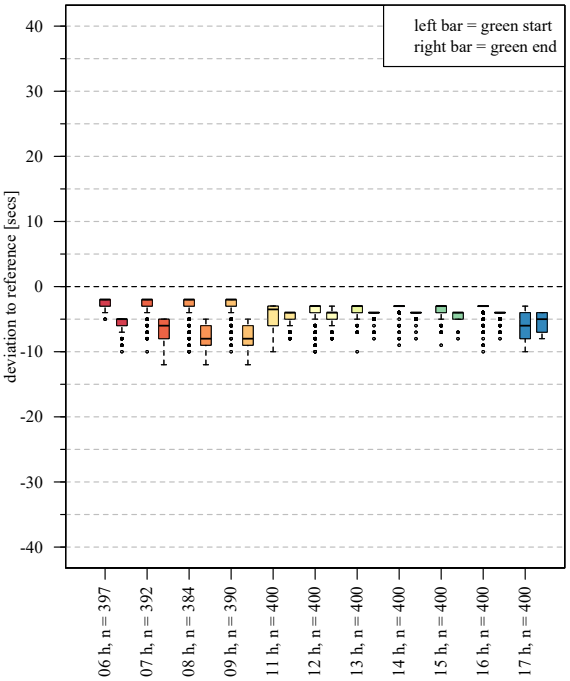
(b) QKZ for $m = 6$



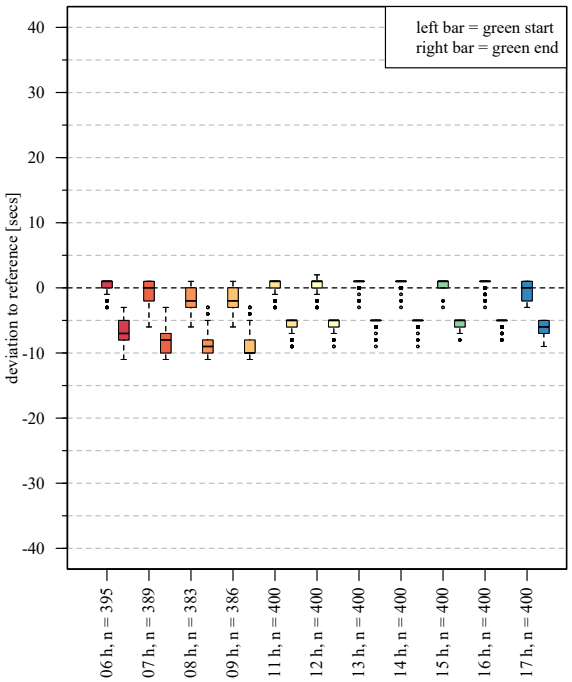
(c) QKZ for $m = 12$



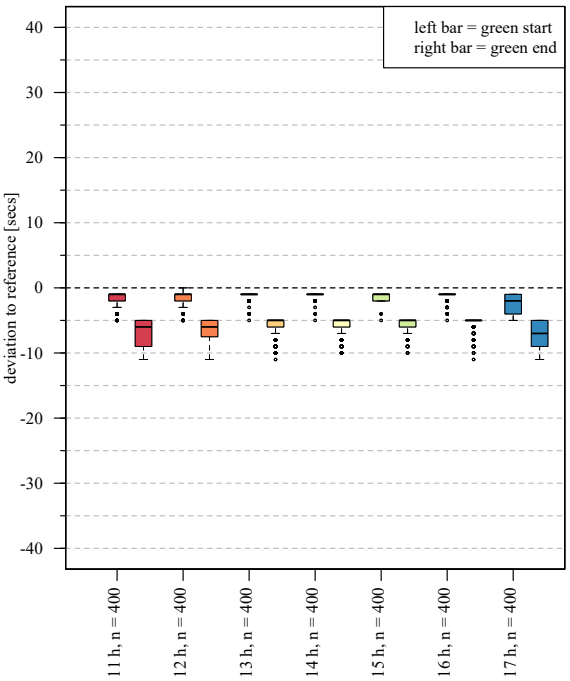
(a) Green accuracy for $m = 1$



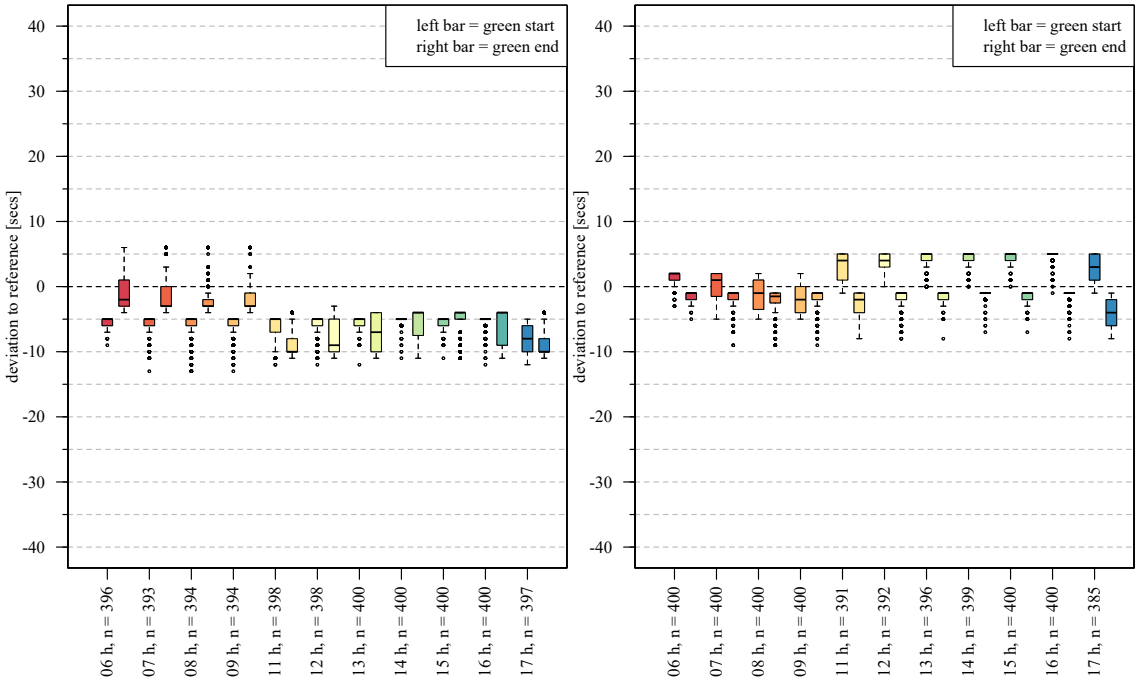
(b) Green accuracy for for $m = 2$



(c) Green accuracy for $m = 5$

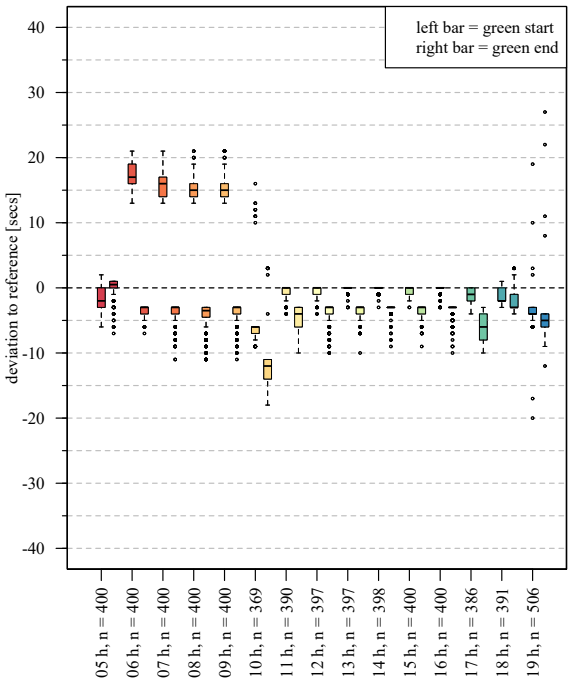


(d) Green accuracy for $m = 6$



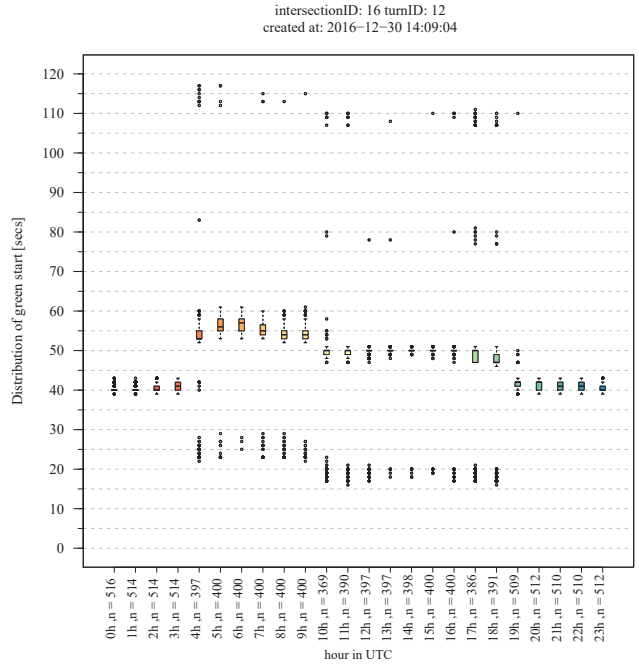
(a) Green accuracy for $m = 8$

(b) Green accuracy for for $m = 11$

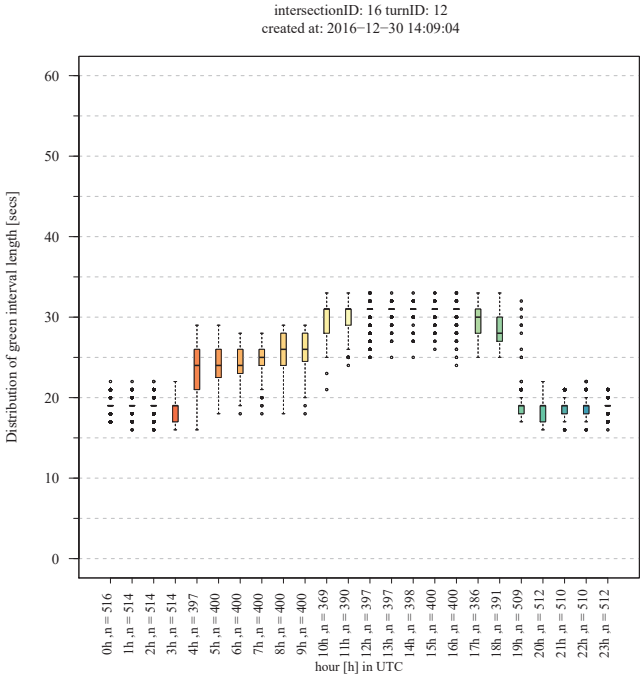


(c) Green accuracy for $m = 12$

D.4.4 Ground Truth Signal Timing



(a) Movement $m = 12$



(b) Movement $m = 12$

Figure D.4.11: Distribution of green starts and green interval duration

Appendix E

Pseudocodes

E.1 Best Cluster Selection

Algorithm 1: Return the best cluster from the cluster set cls

Input: cls
Output: TA
// cls contains a set of clusters $cl_{m,nr}$
// $L()$ returns count of hours of a cluster
// $l()$ returns count of interconnected hours of a cluster
// $h()$ returns corresponding hours of a cluster
$cl_{best} = \emptyset$
$l_{max} = max(l(cls))$
// Maximize count of interconncted hours
foreach $cl_{m,nr} \in cls$ **do**
  if $l(cl_{m,nr}) == l_{max}$ **then**
    $cl_{best} = cl_{best} \cup cl_{m,nr}$
  end
end
if $|cl_{best}| > 1$ **then**
  // Only one best cluster is allowed
  // Maximize count of hours
  $L_{max} = max(L(cl_{best}))$
  foreach $cl_{m,nr} \in cl_{best}$ **do**
    if $L(cl_{m,nr}) == L_{max}$ **then**
      $cl_{best} = cl_{m,nr}$
    end
  end
end
$TA = TA \cup h(cl_{best})$
return TA

E.2 Binary State Vector Substitution Logic

Algorithm 2: Binary state vector substitution logic

Input: $state, const_t$
Output: $modstate$
$peaks = getPeaks(state)$, $gaps = getGaps(state)$
$P = |peaks|$, $G = |gabs|$
$modstate = state$
if $P > 1$ **then**
  // Force while loop once
  $before = 0, after = 1$
  while $(before \neq after) \wedge (P > 1)$ **do**
    // Reassign variable before
    $before = modstate$
    for $p = 1$ *to* P **do**
      // Substitute abundant peaks
      if $(t(gap_p) \geq \delta const_t) \wedge (t(gap_{p+1}) \geq \delta const_t)$ **then**
        $from = peak_p.start, to = peak_p.end$
        $modstate[from \dots to] = [0 \dots 0]$
      end
    end
    // Update values after 1st pass
    $peaks = getPeaks(modstate)$, $gaps = getGaps(modstate)$
    $P = |peaks|$, $G = |gabs|$
    for $g = 1$ *to* G **do**
      // Substitute abundant gaps
      if $(t(peak_g) \geq \delta const_t) \wedge (t(peak_{g+1}) \geq \delta const_t)$ **then**
        $from = gap_g.start, to = gap_g.end$
        $modstate[from \dots to] = [1 \dots 1]$
      end
    end
    // Update values after 2nd pass
    $peaks = getPeaks(modstate)$, $gaps = getGaps(modstate)$
    $P = |peaks|$, $G = |gabs|$
    // Reassign variable after
    $after = modstate$
  end
else
end
return $modstate$

APPENDIX F

Backward calculation methods

The following listing cites literally the backward calculation methods of the OCIT Outstations standard [56] .

1. **Backward calculation - RRV UTC**

   In this simple but not very widespread back calculation method the current UTC time is used as the reference time. The current UTC second modulo of the cycle time of the currently running signal plan must be taken for calculating the current reference second.

2. **Backward calculation - RRV 1.1**

   This widespread back calculation method uses the 00:00:00 local time of 01-01 of the current year as the reference time. The back calculation second (RRS) is calculated from the seconds elapsed since the reference time with regard to the current local time. The daylight-savings-time shift of one hour is considered as elapsed seconds in this method, i.e. the RRS jumps by 3,600 seconds with the daylight-savings-time switch as if the skipped time had actually elapsed. This works the same way with the switch back to standard time.

3. **Backward calculation - RRV**

   This less widespread back calculation method uses 1980-01-01 00:00:00 as a reference date. The back calculation second (RRS) is calculated from the seconds actually elapsed since the reference time. The daylight-savings-time shift of one hour is not taken into account in this method, i.e. the RRS does not jump upon the daylight-savings-time switch but rather continues running steadily.

4. **Backward calculation - RRV midnight**

   This less widespread back calculation method uses midnight (00:00:00) as the reference time. The back calculation second (RRS) is calculated from the seconds elapsed since the reference time with regard to the current

time. The daylight-savings-time shift of one hour is considered as elapsed seconds in this method, i.e. the RRS jumps by 3,600 seconds with the daylight-savingstime switch as if the skipped time had actually elapsed. This works the same way with the switch back to standard time. This is, however, only relevant on the two switch days and has no impact on all other days.

Bibliography

[1] ADAC. Adac stauscanner für iphone & android. bil, ADAC: Allgemeiner Deutscher Automobil-Club, 2016.

[2] Theodore W Anderson. On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, pages 1148–1159, 1962.

[3] Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

[4] Steffen Axer and Bernhard Friedrich. Level of service estimation based on low-frequency floating car data. *Transportation Research Procedia*, 3:1051–1058, 2014.

[5] Steffen Axer and Bernhard Friedrich. Estimating signal phase and timing for traffic actuated intersections based on low frequency floating car data. In *19th International Conference on Intelligent Transportation Systems*, Rio de Janeiro, 2016.

[6] Steffen Axer and Bernhard Friedrich. A methodology for signal timing estimation based on low frequency floating car data: Analysis of needed sample sizes and influencing factors. In *7th International Symposium on Highway Capacity and Quality of Service*, Berlin, Germany, 2016.

[7] Steffen Axer and Bernhard Friedrich. Signal timing estimation based on low frequency floating car data. In *14th World Conference on Transport Research*, Shanghai, China, 2016.

[8] Steffen Axer, Federico Pascucci, and Bernhard Friedrich. Estimation of traffic signal timing data and total delay for urban intersections based on low frequency floating car data. In *Mobil.Tum 2015*, München, Germany, 2015.

[9] Steffen Axer, Jannis Rohde, and Bernhard Friedrich. Level of service estimation at traffic signals based on innovative traffic data services and collection techniques. *Procedia-Social and Behavioral Sciences*, 54:159–168, 2012.

[10] Mirko Barthauer and Bernhard Friedrich. Evaluation of a signal state prediction algorithm for car to infrastructure applications. *Transportation Research Procedia*, 3M:982–991, 2014.

[11] Norman Biggs, E Keith Lloyd, and Robin J Wilson. *Graph Theory, 1736-1936*. Oxford University Press, 1976.

[12] David E Bloom, David Canning, and Günther Fink. Urbanization and the wealth of nations. *Science*, 319(5864):772–775, 2008.

[13] Klaus Bogenberger and Simone Weikl. Quality management methods for real-time traffic information. *Procedia-Social and Behavioral Sciences*, 54:936–945, 2012.

[14] Robert Braun, Fritz Busch, Carsten Kemper, Robert Hildebrandt, Florian Weichenmeier, Cornelius Menig, Ingrid Paulus, and Renate Preßlein-Lehle. TRAVOLUTION - Netzweite Optimierung der Lichtsignalsteuerung und LSA-Fahrzeug-Kommunikation. *Straßenverkehrstechnik*, 53:365–374, 2009.

[15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[16] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[17] Greg Bremser. Coordinating traffic signals for field technicians. *IMSA Journal*, 1(4), 2013.

[18] Michael W Browne. Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132, 2000.

[19] Bi Yu Chen, Hui Yuan, Qingquan Li, William HK Lam, Shih-Lung Shaw, and Ke Yan. Map-matching algorithm for large-scale low-frequency floating car data. *International Journal of Geographical Information Science*, 28(1):22–38, 2014.

[20] Arnaud Cordier, Rémi Domingues, Anthony Labaere, Nicolas Noel, Adrien Thiery, Thomas Cerqueus, Siobhán Clarke, Pawel Idziak, Hui Song, Philip Perry, et al. Dynamic adaptation of the traffic management system cardemo. In *2014 IEEE Eighth International Conference on Self-Adaptive and Self-Organizing Systems*, pages 193–194. IEEE, 2014.

[21] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.

[22] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[23] Hans-Friedrich Eckey, Reinhold Kosfeld, and Christian Dreger. *Statistik: Grundlagen-Methoden-Beispiele*. Springer-Verlag, 2013.

[24] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[25] Julian Faraway, George Marsaglia, John Marsaglia, and Adrian Baddeley. *goftest: Classical Goodness-of-Fit Tests for Univariate Distributions*, 2015. R package version 1.0-3.

[26] S Alireza Fayazi, Ardalan Vahidi, Grant Mahler, and Andreas Winckler. Traffic signal phase and timing estimation from low-frequency transit bus data. *Intelligent Transportation Systems, IEEE Transactions on*, 16(1):19–28, 2015.

[27] FGSV. *Hinweise zur mikroskopischen Verkehrsflusssimulation: Grundlagen und Anwendungen*. Fgsv. Nr: 388. FGSV-Verl., Köln, ausg. 2006 edition, 2006.

[28] FGSV. *Richtlinien für Lichtsignalanlagen, RiLSA - Lichtzeichenanlagen für den Straßenverkehr*. Fgsv. Nr: 321. FGSV-Verl., Köln, ausg. 2010 edition, 2015.

[29] Bernhard Friedrich, Markus Friedrich, and Christian Priemer. Impacts of the service quality of single road facilities on the service quality in networks. In *TRB 5th International Symposium on Highway Capacity and Quality of Service*, pages 381–390, 2006.

[30] Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R Sveinsson. Random forest classification of multisource remote sensing and geographic data. In *Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. Proceedings. 2004 IEEE International*, volume 2, pages 1049–1052. IEEE, 2004.

[31] Google. *Google Earth*, December 2016.

[32] Philippe Grosjean and Frederic Ibanez. *pastecs: Package for Analysis of Space-Time Ecological Series*, 2014. R package version 1.3-18.

[33] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.

[34] James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[35] Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[36] Martin Hauschild. Weather and road surface information and hazard warnings data content acquisition through advanced probe vehicle systems. In *Proc. 12th World Congress Intelligent Transport Systems*, 2005.

[37] Timothy Hunter, Pieter Abbeel, and Alexandre M Bayen. The path inference filter: model-based low-latency map matching of probe vehicle data. In *Algorithmic Foundations of Robotics X*, pages 591–607. Springer, 2013.

[38] Daniel Jiang and Luca Delgrossi. IEEE 802.11 p: Towards an international standard for wireless access in vehicular environments. In *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*, pages 2036–2040. IEEE, 2008.

[39] Konstantinos Katsaros, Ralf Kernchen, Mehrdad Dianati, David Rieck, and Charalambos Zinoviou. Application of vehicular communications for improving the efficiency of traffic in urban areas. *Wireless Communications and Mobile Computing*, 11(12):1657–1667, 2011.

[40] Markus Kerper, Christian Wewetzer, Andreas Sasse, and Martin Mauve. Learning traffic light phase schedules from velocity profiles in the cloud. In *New Technologies, Mobility and Security (NTMS), 2012 5th International Conference on*, pages 1–5. IEEE, 2012.

[41] Peter Koonce, Lee Rodegerdts, Kevin Lee, Shaun Quayle, Scott Beaird, Cade Braud, Jim Bonneson, Phil Tarnoff, and Tom Urbanik. Traffic signal timing manual. Technical report, Federal Highway Administration, 2008.

[42] Emmanouil Koukoumidis, Li-Shiuan Peh, and Margaret Rose Martonosi. Signalguru: leveraging mobile phones for collaborative traffic signal schedule advisory. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, pages 127–140. ACM, 2011.

[43] Paul Krijger. *Traffic Light Prediction for Tom Tom Devices*. Eindhoven University of Technology, 2013.

[44] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[45] Jurgen Lohrer and Markus Lienkamp. Building representative velocity profiles using fastdtw and spectral clustering. In *ITS Telecommunications (ITST), 2015 14th International Conference on*, pages 45–49. IEEE, 2015.

[46] Yin Lou, Chengyang Zhang, Yu Zheng, Xing Xie, Wei Wang, and Yan Huang. Map-matching for low-sampling-rate GPS trajectories. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 352–361, New York, NY, USA, 2009. ACM.

[47] George Marsaglia, John Marsaglia, et al. Evaluating the anderson-darling distribution. *Journal of Statistical Software*, 9(2):1–5, 2004.

[48] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[49] Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[50] William Mendenhall, Robert J Beaver, and Barbara M Beaver. *Introduction to probability and statistics*. Cengage Learning, 2012.

[51] Stadt München. *Verkehrsmengenkarte*. Landeshauptstadt München, 2009.

[52] Marko Modsching, Ronny Kramer, and Klaus ten Hagen. Field trial on gps accuracy in a medium size city: The influence of built-up. In *3Rd workshop on positioning, navigation and communication*, pages 209–218, 2006.

[53] Carsten Moeller. *OSM2PO - OpenStreetMap converter and routing engine for Java*, 2016.

[54] David J Murray-Smith. *Testing and Validation of Computer Simulation Models*. Springer, 2015.

[55] Thorsten Neumann. *Rückstaulängenschätzung an Lichtsignalanlagen mit Floating-Car-Daten*. PhD thesis, Institut für Verkehrssystemtechnik, Deutsches Zentrum für Luft- und Raumfahrt, 2011.

[56] ODG. Ocit outstations traffic signal controllers. Technical report, OCIT Developer Group, 2012.

[57] R Lyman Ott and Micheal T Longnecker. *An introduction to statistical methods and data analysis*. Cengage Learning, 2008.

[58] Markos Papageorgiou, Christina Diakaki, Vaya Dinopoulou, Apostolos Kotsialos, and Yibing Wang. Review of road traffic control strategies. *Proceedings of the IEEE*, 91(12):2043–2067, 2003.

[59] Dieter Pfoser and Christian S Jensen. Capturing the uncertainty of moving-object representations. In *International Symposium on Spatial Databases*, pages 111–131. Springer, 1999.

[60] T. Pohlmann. *New Approaches for Online Control of Urban Traffic Signal Systems*. PhD thesis, Institut für Verkehr und Stadtbauwesen, TU-Braunschweig, 2011.

[61] Christian Priemer. *Kommunikationsdatenbasierte, dezentrale Lichtsignalsteuerung in städtischen Netzen*. Shaker Verl, 2011.

[62] Valentin Protschky, Stefan Feit, and Claudia Linnhoff-Popien. On the potential of floating car data for traffic light signal reconstruction. In *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, 2015.

[63] Valentin Protschky, Christian Ruhhammer, and Stefan Feit. Learning traffic light parameters with floating car data. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 2438–2443. IEEE, 2015.

[64] Valentin Protschky, Paul Seifert, and Stefan Feit. Stop line detection using satellite-image segmentation. In *IEEE 81st Vehicular Technology Conference (VTC-Spring 2015)*, pages 1–5, 2015.

[65] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[66] Mahmood Rahmani and Haris N Koutsopoulos. Path inference from sparse floating car data for urban networks. *Transportation Research Part C: Emerging Technologies*, 30:41–54, 2013.

[67] Paul Ramsey et al. PostGIS manual. *Refractions Research Inc*, 2005.

[68] Felix Rempe, Philipp Franeck, Ulrich Fastenrath, and Klaus Bogenberger. Online freeway traffic estimation with real floating car data. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 1838–1843. IEEE, 2016.

[69] Hans-Georg Retzko and Manfred Boltze. Timing of intergreen periods at signalized intersections: The german method. *ITE journal*, 7(1):23–26, 1987.

[70] Ralf-Peter Schäfer, Astrid Gühnemann, and Kai-Uwe Thiessenhusen. Neue ansätze im verkehrsmonitoring durch floating car daten. In *19. Verkehrswissenschaftliche Tage (CD-ROM)*, 2003.

[71] Johannes Schlaich, Thomas Otterstätter, and Markus Friedrich. Generating trajectories from mobile phone data. In *Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies*, 2010.

[72] Robert Schönauer, Gerald Richter, Markus Straub, Christian Rudloff, and Melitta Dragaschnigg. Red or green: Estimating the patterns of traffic signal through cyclists' gps tracks for real time navigation. In *Transportation Research Board 93rd Annual Meeting*, volume 14, page 1664, 2014.

[73] Statista - das Statistik-Portal: Statistiken, Marktdaten & Studien. *Anteil der in Städten lebenden Bevölkerung von 1950 bis 2030 in Deutschland und weltweit*, 2017.

[74] Michael A Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347):730–737, 1974.

[75] W Richard Stevens and Stephen A Rago. *Advanced programming in the UNIX environment*. Addison-Wesley, 2013.

[76] Le Hung Tran, Quoc Viet Hung Nguyen, Ngoc Hoan Do, and Zhixian Yan. Robust and Hierarchical Stop Discovery in Sparse and Diverse Trajectories. Technical report, EPFL, 2011.

[77] Thomas Wietholt. *Einsatzbereiche Grüner Wellen und verkehrsabhängiger Steuerungen*. PhD thesis, Uni Bochum, 2008.

[78] Daniel Wollschläger. *R kompakt: Der schnelle Einstieg in die Datenanalyse*. Springer-Verlag, 2016.